Méthodes d'apprentissage et approches expérimentales appliqués aux réseaux d'interfaces protéiques

Mounia Achoch

UNIVERSITÉ GRENOBLE ALPES

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ GRENOBLE ALPES

Spécialité : **Doctorat Science Pour l'Ingénieur**

Arrêté ministériel : 7 août 2006

Présentée par

Mounia ACHOCH

Thèse dirigée par **Mr Kavé SALAMATIAN**
Codirigée par **Mme Claire LESIEUR**

Préparée au sein du Laboratoire **LISTIC**

dans **les Écoles Doctorales SISEO**

Méthodes d'apprentissage et approches expérimentales appliqués aux réseaux d'interfaces protéiques

Thèse soutenue publiquement le **« 30 Septembre 2015 »**,
devant le jury composé de :

Mme Frédérique LISACEK
Swiss Institute of Bioinformatics, Rapporteur
Mr Alexandre DE BREVERN
Directeur de recherche, Université de Paris Diderot, Rapporteur
Mme Sylvie RICARD BLUM
Professeur, Université Claude Bernard Lyon 1, Examinateur
Mr Rachid JALAL
Professeur, Faculté des sciences et technique Maroc, Examinateur
Mr Laurent VUILLON
Professeur, Université Savoie Mont Blanc France, Président du jury
Mme Claire LESIEUR
Chargé de recherche, CNRS, co-directeur de thèse
Mr Kavé SALAMATIAN
Professeur, Université Savoie Mont Blanc France, directeur de thèse

Dédicace spéciale

Je dédie ce modeste travail à :

Mon cher père, je me rappelle toujours de tous les moments où tu m'as poussé à travailler et à réussir, j'avoue que si je suis devenue quelque chose actuellement c'est grâce à tes efforts, à tes conseilles et à ta surveillance.

Ma très chère mère, j'aimerai toujours te remercier pour tous ce que tu as fait jusqu'à notre jours-là pour assurer l'éducation et la formation de tous tes enfant. J'avoue vraiment que tu été pour moi la lumière qui me guide mes routes et qui m'emmène aux chemins de la réussite, c'est grâce à toi que je dois toute ma réussite.

Mon très cher mari, aucun mot ne saurait t'exprimer mon profond attachement et ma reconnaissance pour l'amour, la tendresse et la gentillesse dont tu m'as toujours entouré. J'aimerai bien que tu trouves dans ce travail l'expression de mes sentiments de reconnaissance les plus sincères car grâce à ton aide et à ta patience avec moi que ce travail a pu voir le jour.

Ma chère sœur et mon cher frère, J'espère que mon travail sera le témoignage de mon respect et de mes sentiments les plus sincères.

A toute la famille et à toutes les personnes qui ont servis pour ma formation, mon éducation et mon enseignement.

Mounia ACHOCH

À mon cher frère
À ma chère sœur
À mon très cher Soufiane
À mes chers amis

Remerciements

J'aimerais tout d'abord adresser mes profonds respects à Monsieur Patrick LAMBERT directeur du Laboratoire d'Informatique, Systèmes, Traitement de l'Information et de la Connaissance (LISTIC) et Madame Sylvie GALICHET (ancienne directrice), pour m'avoir offert l'occasion de passer ma thèse au sein de leur laboratoire.

Un grand merci à Monsieur Kavé SALAMATIAN pour son accueil, je lui suis très reconnaissant de m'avoir laissée une grande liberté pour exploiter mes idées et Madame Claire LESIEUR pour son aide et l'attention dont elle m'a entourée tout au long de ce travail. Ainsi pour leurs remarques, leurs suggestions pertinentes, leurs encouragements, leur précieuse direction et leur confiance qui m'ont énormément servi et permis d'apprendre rapidement, je les remercie.

Je tiens à remercier la communauté de recherche académique (ARC6) qui a financé ma thèse. Je remercie aussi tous les collègues qui m'ont aidée de près ou de loin au cours de mes trois ans de thèse, notamment l'ensemble des collaborateurs de LISTIC, pour leur bonne humeur permanente et leur chaleureux accueil.

Je présente mes remerciements les plus sincères à Monsieur Laurent VUILLON, professeur au Laboratoire d'Analyse et de Mathématiques Appliquées (LAMA) et Monsieur Rodrigo Dorantes-Gilardi, doctorant au LAMA, pour cette expérience d'un travail en groupe et l'aide qu'ils ont pu me fournir quand j'en avais besoin.

Je tiens à exprimer ma gratitude à Madame Frédérique LISACEK et Monsieur Alexandre DE BREVERN de m'avoir fait l'honneur d'accepter d'être rapporteur de ce travail.

Toute ma reconnaissance va également à Madame Sylvie RICARD BLUM et Monsieur Rachid JALAL pour l'honneur qu'ils m'ont fait d'accepter de participer à mon jury.

Cette thèse n'aurait sûrement pas été la même sans les diverses collaborations. Dans ce sens, je remercie Monsieur Giovanni FEVERATI, qui m'a lancé sur le programme GEMINI.

Résumé

Cette étude s'inscrit dans le cadre d'un problème biologique et son objectif est de comprendre les mécanismes d'assemblage des protéines. L'assemblage d'une protéine en oligomère est particulièrement important car il est impliqué dans de nombreuses pathologies allant de l'infection bactérienne aux maladies de type Alzheimer ou même des cancers. L'assemblage protéique est un mécanisme de combinaison de deux ou plusieurs chaînes protéiques, il est aussi par ailleurs souvent utilisé par les organismes vivants pour déclencher une activité biologique. La sous unité B de la toxine du choléra ($CtxB_5$), qui appartient à la famille des toxines AB_5, est étudiée comme modèle principal de l'assemblage. Des résultats expérimentaux ont fourni des informations sur l'assemblage de la toxine mettant en avant l'implication de certains acides aminés. La première question que j'ai abordée pendant ma thèse était de comprendre leur rôle et de voir si les approches réseaux étaient pertinentes pour y répondre. J'ai pu montrer en utilisant des mutations d'acides aminés que ces derniers s'influençaient entre eux suivant des mécanismes en cascade ou de « Peer to Peer » afin de coordonner les étapes de l'assemblage (les chapitres 4, 5 et 6). La structure et la fonction des protéines sont définies par des séquences d'acides aminés qui varient naturellement en raison de mutation génétique. J'ai donc décidé d'élargir ce champ d'investigation pour voir si le mécanisme en cascade était généralisable comme moyen de perturber une structure de protéine par le biais d'une mutation. Ici il s'agit de comprendre les changements de structure liés à des mutations et pouvant menés à des maladies. J'ai tout d'abord étudié des jeux de données pour connaître les caractéristiques réseaux de protéines saines (chapitre 7, 8 et 9), avant de regarder l'effet de la mutation systématique de chacun des acides aminés de $CtxB_5$ sur sa structure globale (chapitre 10 et 11). Les mutations peuvent engendrer des changements de structure modérés ou très grand autour de l'acide aminé muté ou à des distances très éloignées. Ces résultats sont consistants avec tous les effets connus de mutation : robustesse (maintien de la fonction), évolution ou adaptation (émergence d'une nouvelle fonction) et fragilité (pathologies). Les résultats montrent aussi une faible corrélation entre le nombre de contacts d'un acide aminé et la quantité de changement structuraux induit par sa mutation. Il n'est donc pas simple d'anticiper l'effet d'une mutation : Le dernier chapitre de ma thèse aborde ce problème (chapitre 12).

Mots clés : protéine, assemblage, réseaux, interface

Abstract

The aim of this study is to understand protein assembly mechanisms. The assembly of a protein in an oligomer is particularly important because it is involved in many pathologies going from bacterial infection, Alzheimer like diseases or even some cancers. Protein assembly is the combination of two or more protein chains to induce a biological activity. The B subunit of the cholera toxin pentamer ($CtxB_5$), which belongs to the family of AB_5 toxins, is studied as the main model of assembly. Experimental results have provided information on the assembly of the toxin highlighting the involvement of certain amino acids. The first problem addressed in my thesis is to understand their role and see if network approaches are relevant to such investigation. I was able to show using amino acid mutations, that amino acids influence each other by cascade or "peer to peer" mechanisms in order to coordinate the various steps of the assembly (Chapters 4, 5 and 6). The structure and function of the proteins are defined by amino acid sequences which naturally vary due to genetic mutation. So I decided to expand this field of investigation to see if the cascade mechanism was generalized as a mean of disrupting a protein structure. Here it is to understand how a protein loses its function by way of a significant change of structure upon mutation. First, I studied dataset to know the characteristics of healthy protein networks (Chapter 7, 8 and 9), and after I looked at the effects of the systematic mutation of each amino acid of $CtxB_5$ on its overall structure (Chapter 10 and 11). Mutations led from moderate to very large structural changes around the mutated amino acid or at long distances. These results are consistent with known effects of mutation: robustness (maintenance function), evolution or adaptation (emergence of a new feature) and fragility (pathologies). The results also show a weak correlation between the number of amino acid contacts of the mutated amino acid and the amount of structural change induced by its mutation. It is therefore not easy to anticipate the effect of a mutation: The last chapter of my thesis addresses this problem (Chapter 12).

Key words: protein, assembly, networks, interfaces.

Table des matières

Introduction générale

Le terme protéine vient du grec « proteios » qui signifie premier en importance [1]. Ce terme reflète l'omniprésence de ces molécules dans les êtres vivants. Que ce soient les bactéries, les plantes ou les animaux, tous sont essentiellement constitués d'eau et de protéines. Ces dernières interviennent à tous les stades du fonctionnement de notre organisme, par exemple l'hémoglobine transporte l'oxygène, l'insuline régule le taux de sucre, les anticorps combattent les infections, la myosine permet à nos muscles de se contracter et les collagènes constituent nos tendons et nos ligaments.

Les protéines sont synthétisées par une machinerie complexe, le ribosome. A partir de l'information génétique codée dans l'ADN en passant par l'ARN, le ribosome construit une chaîne linéaire d'acides aminés qui adopte une structure tridimensionnelle répondant à sa fonction biologique. Les protéines sont des bio-polymères naturels avec un squelette très simple, dont la diversité physicochimique est portée par la chaîne latérale variable des 22 acides aminés. Cette diversité d'éléments de bases produit une large variété de structures (chapitre 1). La structure plus communément appelée la forme d'une protéine est très importante puisqu'elle supporte la fonction et donc l'activité biologique de la protéine. Une illustration est la protéine dit désordonnée c'est-à-dire pas de structure 2D/3Ddont une forme adapté à leur flexibilité ponctuelle.

Les fonctions biologiques des protéines sont diverses, d'enzyme à des rôles purement structuraux. La structure fonctionnelle des protéines est appelée forme native. Le processus d'acquisition de la structure native est le repliement de la protéine. Il existe des pathologies liées à de mauvais repliements qui mènent à des changements de formes tels que la fonction biologique est perdue. Ces mauvais repliements peuvent naître des perturbations internes telles que des mutations ponctuelles ou externes comme des modifications dans l'environnement. Aussi, la compréhension fine de la structure des protéines, en dehors du problème fondamental sous-jacent, revêt une importance majeure dans le cadre de la recherche médicale.

L'association de chaînes protéiques entre elles ou avec des cofacteurs/ligands est un mécanisme souvent utilisé par les organismes vivants pour déclencher une activité biologique.

Dans le présent travail, seules les associations permanentes entre chaînes protéique ont été étudiées. Les associations transitoires ou avec des ligands ne sont pas traitées. Les dimères, association de deux chaines protéiques ne sont pas non plus considérés, et seuls les cas d'association d'au moins trois chaines sont étudiés. Dans ces cas-là, chaque chaîne individuelle fournit un ou des domaines qui s'associent avec un ou des domaines fournis par les autres chaînes pour former une interface protéique et ainsi construire une protéine oligomérique (polymère/oligomère) (Figure 1.1). L'assemblage protéique implique deux réactions: le repliement de la chaîne individuelle (interactions intramoléculaires) et l'association des différentes chaînes (interactions intermoléculaires). Les contacts intermoléculaires constituent l'interface protéique.

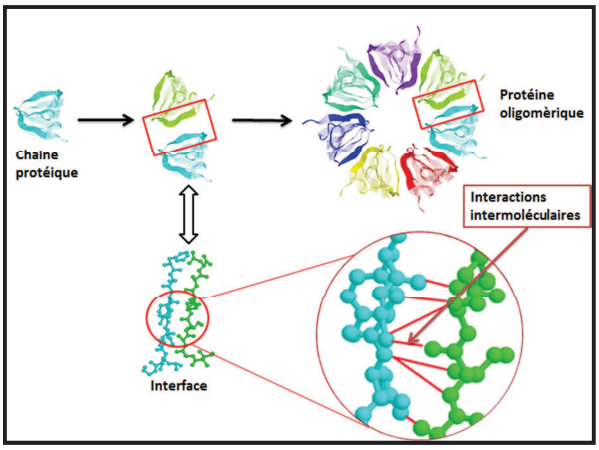


Figure 1.1 : Exemple d'interface entre deux chaînes protéiques lors du mécanisme d'assemblage. L'heptamère de la co-chaperone 10 de *Mycobacterium tuberculosis* est utilisé comme illustration (PDB 1HX5).

Il existe aussi des maladies liées à des changements locaux, à savoir des mutations ponctuelles d'acides aminés menant à une dissociation de l'oligomère et sa réassociation en une forme alternative non fonctionnelle. Ce type de perte de fonction biologique mène à une maladie, dite maladie conformationnelle, telles que les maladies d'Alzheimer, de Parkinson ou encore certains cancers impliquant la protéine p53, un suppresseur de tumeurs [2].

Des oligomères de protéines sont aussi impliquées dans des maladies comme facteurs de virulence, par exemple, la toxine cholérique responsable de la maladie du choléra. Il est essentiel de comprendre comment ces protéines s'assemblent en oligomères pour prédire leur assemblage et concevoir des inhibiteurs capables de le prévenir. Plus exactement, la conception de tels inhibiteurs implique d'identifier les acides aminés clés dans la formation d'une interface et de comprendre leur rôle dans cette formation.

Il existe pour le moment quatre grandes classes de protéines : globulaires, fibrillaires, transmembranaires et désordonnées. Ces protéines peuvent être groupées en trois catégories géométriques très probablement distinctes en termes de « densités » de contacts (nombre d'interactions atomiques par volume ou surface). Les protéines globulaires sont très compactes, elles maximisent le nombre d'interactions et minimisent les « vides ». Les protéines globulaires peuvent être décrites de façon approximative par une sphère. Les protéines fibrillaires sont de longues molécules en forme de filaments. Elles favorisent les interactions le long d'une direction en particulier, en première approximation leurs formes peuvent être décrites par un cylindre. Les protéines désordonnées restent les moins décrites, leurs structures sont peu ordonnées, peu rigides et très flexibles suggérant qu'elles reposent sur moins d'interactions que les protéines globulaires ou fibrillaires. Mon travail de thèse s'intéresse aux protéines oligomériques globulaires et en particulier sur la formation de leurs interfaces entre monomères.

Des biophysiciens déterminent les structures atomiques par des mesures de diffraction aux rayons X sur une protéine cristallisée et par la résonance magnétique nucléaire (RMN) pour des protéines en solution. Du fait des difficultés techniques, du coût et du temps que demandent ce type d'expérimentation, la grande majorité des structures restent encore inconnues.

Les théoriciens participent également à cet effort en proposant des modèles et des méthodes de simulation qui permettent d'accéder à la dynamique des protéines. Pour simuler des systèmes biologiques, deux grandes techniques ont été développées. Premièrement, la dynamique moléculaire qui est une méthode qui tente de reproduire le comportement structurale d'une protéine au cours du temps. Deuxièmement, l'approche réseau qui modélise une structure atomique sous la forme d'un graphe d'atomes ou d'acides aminés en interaction dont la dynamique, l'architecture et les propriétés servent à comprendre les structures. Par exemple, cette approche permet d'analyser une interface et d'identifier les acides aminés qui la composent. Elle peut aussi s'avérer utile pour modéliser un réseau sous-jacent une protéine désordonnée [3].

Cette approche graphe, dite approche de type réseau, est maintenant bien reconnue [4, 5]. La dynamique moléculaire et l'approche réseau peuvent être combinées pour analyser des changements de structures [6-8].

Mon sujet de thèse consiste à étudier les propriétés structurales des protéines oligomériques (Chapitre 1) par des approches réseaux (chapitre 2). L'idée étant de modéliser une structure atomique de protéine par un réseau d'interactions entre atomes, interactions définies en fonction des distances euclidiennes entre atomes. Cette modélisation s'applique donc uniquement aux protéines de structures atomiques connues ou à des protéines ayant des structures homologues. J'utilise des notions et des propriétés connues de réseaux pour tester des hypothèses sur le mécanisme de l'assemblage protéique comme la coordination entre les interactions intramoléculaires et intermoléculaires, modélisées par deux réseaux différents ou encore l'effet de mutations pour proposer quels résidus serait susceptible de provoquer une dissociation des chaînes d'un oligomère.

En général les approches expérimentales sur les protéines sont chères, empiriques, et laborieuses, par exemple lorsqu'il s'agit de distinguer les acides aminés nécessaires et suffisants à la formation des interfaces. Un soutien par des approches bioinformatiques est intéressant puisqu'il permet de prédire ses résidus, appelés souvent 'hots spots' ou points chauds, réduisant ainsi le champ d'investigation expérimental. De nombreux algorithmes capables d'identifier les points chauds à partir de la structure atomique d'oligomères sont disponibles et accessible via différentes bases de données dans lesquelles les interfaces ont été répertoriées, comme SCOPPI, SCOWLP, PPI, PIBASE, PRISM, etc. [9-11].

Les coordonnées atomiques sont nécessaires pour l'utilisation de ses algorithmes et elles sont accessibles auprès de la Protein DataBank (www.rcsb.org/pdb/). Ces algorithmes fournissent la matière première nécessaire pour aborder la question des mécanismes de formation des interfaces par des méthodes théoriques, en particulier du fait qu'ils ouvrent les portes à des analyses sur des grands jeux de données. J'ai donc suivi une approche similaire en étudiant les propriétés de grands jeux de données sur les interfaces par des algorithmes ad-hoc ou en étudiant un prototype unique, le pentamère de la sous-unité B de la toxine du choléra ($CtxB_5$) afin de déterminer les paramètres importants pour l'interface. Dans tous les cas, j'utilisais un algorithme pour identifier des « hots spots » dans un oligomère et je modélisais l'interface protéique par un graphe avec comme nœuds les hots spots et comme liens, les interactions entre hots pots.

Des notions classiques de réseaux ou de graphes tels que les distances géodésiques, les chemins et la propagation se sont avérées utiles pour répondre à nos questions sur les

interfaces et les assemblages. La première notion de réseau que j'ai utilisé pour comprendre les assemblages est la notion de communication dans un réseau. Les éléments de bases du réseau, les acides aminés ou les atomes dans notre cas, communiquent à différentes échelles (distance géodésique), soit ils ont un lien direct et sont à une distance géodésique de 1 ou ils n'ont pas de liens physiques directs et ils sont à une distance géodésique de plus de 1. Donc, une communication à courte ou longue distance est naturelle dans un réseau, ce qui correspond bien à la situation des protéines où les phénomènes allostériques sont connus. Une telle notion n'existe pas en chimie où une interaction se fait uniquement si les atomes sont suffisamment proches dans l'espace, en d'autres termes un lien physique direct. J'ai exploré ces notions de distances géodésiques pour mettre en évidence des mécanismes permettant de coordonner les réactions de repliement et d'association lors d'un assemblage. J'ai ensuite utilisé cette notion pour décrire la plasticité conformationnelle d'une protéine sous l'effet de perturbation via la mutation individuelle de ces acides aminés.

La deuxième notion de réseau que j'ai utilisé pour comprendre les interfaces est le phénomène de propagation d'information dans un réseau. Ce phénomène explique comment une perturbation locale peut altérer fondamentalement un réseau via une propagation large et à longue distance. La propagation est le phénomène sous-jacent une mutation qui peut elle aussi totalement métamorphoser la conformation d'une protéine. J'ai pu montrer l'existence d'un mécanisme de propagation similaire à celui appelé pair-à-pair dans les réseaux informatiques par exemple. En effet, les mutations peuvent aussi engendrer des changements de structure faibles, modérés ou très grand autour de l'acide aminé muté où à des distances très éloignés. Finalement, j'ai aussi travaillé sur la notion de degré et de hubs pour comprendre quelles propriétés pouvaient fragiliser une protéine lors de mutation.

Le but est de comprendre comment les interactions atomiques s'organisent pour créer une structure et lui conférer à la fois robustesse et plasticité. Le premier chapitre décrit les protéines et leurs structures pour comprendre les mécanismes d'assemblage. Il est essentiellement un rappel des bases de biochimie et il est adressé plus précisément à des lecteurs non biologistes avec dans l'idée de leur permettre de cerner les constituants élémentaires des protéines, les structures et formes des protéines et leur dynamiques. Le deuxième chapitre sert lui d'introduction générale pour des lecteurs non versés aux mondes et notions familières aux réseaux. Il explicite aussi par des illustrations, les utilisations possibles pour comprendre les mécanismes d'assemblage protéique. La méthodologie est expliquée dans le chapitre trois où j'ai décrit les différents méthodes utilisées au cours de ma thèse. Les

résultats sont étudiés en deux grands axes : le premier est la mise en évidence de mécanismes d'influence en cascade comme moyen de coordonner l'assemblage, mais aussi de permettre plasticité et robustesse structurale (Chapitres 4, 5, 6 et 11). Dans les chapitres 7, 8 et 9, j'ai étudié des caractéristiques réseaux de protéines saines comme le degré des nœuds, la connectivité du réseau et son architecture. Dans les chapitres 10 et 11, j'ai regardé l'effet de la mutation de chacun des acides aminés de l'interface du pentamère de la sous-unité B de la toxine du choléra sur sa structure globale à partir de mutations *in silico*. Ces travaux ont permis de montrer que le degré d'un nœud n'est pas un paramètre suffisant pour expliquer la fragilité structurale vis-à-vis d'une mutation et que des changements structuraux notoires ne signifient pas nécessairement perte d'intégrité structurale. Finalement, dans le dernier chapitre de ma thèse (chapitre 12), je me suis intéressée à la tolérance structurale d'une protéine vis-à-vis de la mutation en considérant certaines positions dans $CtxB_5$.

Chapitre 1: Assemblage des protéines

1.1 Protéines

Les protéines ont été découvertes en 1839 par Mulder [12]. Elles jouent des rôles cruciaux dans tous les processus biologiques. Les protéines sont formées à partir d'un répertoire de 22 acides aminés distincts qui constituent les éléments de base des protéines.

1.1.1 De l'ADN aux protéines

Les protéines sont des chaînes d'acides aminés dont l'association est gouvernée par le code génétique. L'acide désoxyribonucléique (ADN) est une molécule, en forme de double hélice, qui contient l'information génétique. Cette information est codée avec 4 bases différentes - adénine (A), cytosine (C), guanine (G) et thymine (T) - regroupées en triplets appelés codons. Le code génétique traduit l'information génétique en protéine via l'acide ribonucléique (ARN) (Figure 1.2).

L'expression d'un gène est ainsi constituée de deux étapes. D'une part, la transcription qui est la copie d'une partie de l'information de l'ADN en ARN, dans lequel la base thymine est remplacée par l'uracile (U). D'autre part, la traduction de l'information génétique qui intervient dans les ribosomes où l'enchaînement de codons de l'ARN est converti en acides aminés.

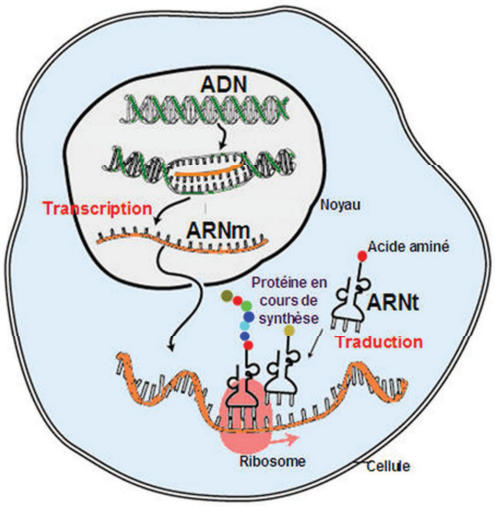


Figure 1.2 : de l'ADN aux protéines.

Parmi les 64 codons possibles, seuls 61 codent pour les 20 acides aminés naturels, les trois codons restants étant des codons de fin de traduction. Plusieurs codons différents représentent donc un même acide aminé, on parle de dégénérescence du code génétique.

Le terme mutation désigne le changement de la séquence des nucléotides dans l'ADN. L'évolution biologique est l'ensemble des modifications que subissent les espèces vivantes, par exemple suite à des mutations au cours du temps. Les conséquences de ces mutations sur la séquence d'acides aminés des protéines sont variables. Les conséquences se mesurent principalement en termes d'impact sur la fonction biologique. La fonction peut être conservée, il s'agit alors de robustesse à la mutation. Les mutations sont fonctionnellement neutres, la présence ou non de changements structuraux est rarement établi. La protéine perd sa fonction sous l'effet d'une mutation, la protéine est sensible ou susceptible à la mutation. En général, les cas étudiés sont associés à des pathologies. Lorsqu'une mutation ou un ensemble de mutation permet l'émergence d'une nouvelle fonction, il s'agit d'*adaptation*. L'adaptation peut naître de mutations individuellement neutres qui combinées avec d'autres et permettent une nouvelle fonction (épistasies). Dans ce cas, on les appelle des mutations adaptatives. Elles se distinguent des mutations neutres car elles ont un potentiel latent d'adaptation, ce phénomène est appelé 'neutral mutational drift' [13]. Bien que cela ne soit pas déterminé systématiquement, les mutations adaptatives induisent nécessairement des changements structuraux neutres fonctionnellement qui sont responsables de l'effet distinct de la deuxième mutation sur leurs formes mutées et la forme de la protéine d'origine (sans mutation).

Les mutations sont le moteur de l'évolution et la source de la diversité entre individus. L'étude de l'effet structural de mutations va me permettre d'aborder la question de la plasticité structurale des protéines. En me concentrant uniquement sur les conséquences structurales des mutations, j'espère voir se dessiner un ensemble de solutions permettant d'encadrer les mécanismes de robustesse, fragilité et adaptation.

1.1.2 Synthèse des protéines

Après une traduction de l'ARN, synthétisée dans les ribosomes, une protéine est obtenue sous la forme d'un assemblage linéaire d'acides aminés. Une protéine a au minimum 50 acides aminés et peut aller jusqu'à plus de 1000 (la protéine titine a 30000 acides aminés).

En moyenne les protéines ont autour de 300 acides aminés. Une chaîne de cinquante acides aminés ou moins s'appelle un peptide.

Les acides aminés d'une protéine sont connectés par les liaisons covalentes appelées liaisons peptidiques [14] (Figure 1.3). Une liaison peptidique se forme entre deux acides aminés quand l'hydrogène du groupe amine de l'un réagit avec l'hydroxyle du groupe carboxylique de l'autre, pour former une molécule d'eau (réaction de condensation). Le groupe/lien CONH s'appelle un groupe peptidique.

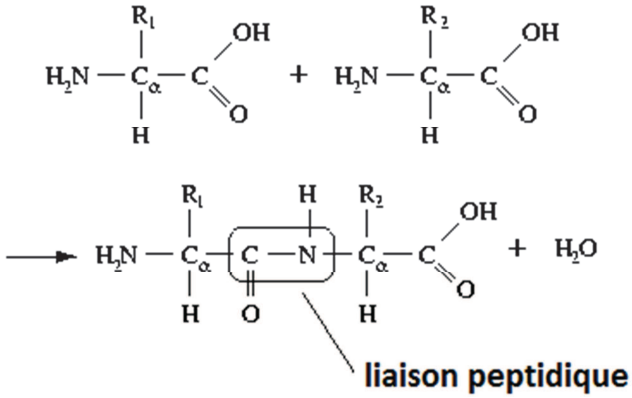


Figure 1.3 : Formation de la liaison peptidique.

Le groupe amine qui n'est pas engagé dans une liaison peptidique est le N-terminal ou N-ter de la protéine, le groupe carboxylique à l'autre extrémité de la protéine est le C-terminale ou C-ter. Par convention, la séquence d'une protéine est écrite de gauche à droite en commençant par le N-terminal (Figure 1.4).

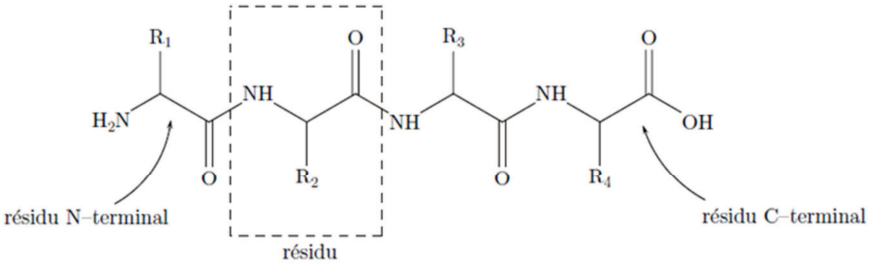


Figure 1.4 : Résidus d'un tétra-peptide avec les extrémités N-terminale et C-terminale [15].

Dans les conditions physiologiques (milieu aqueux et pH égal 7), un peptide adopte spontanément une forme zwitterionique pour laquelle les extrémités N-ter et C-ter sont

ionisées, respectivement en $-NH^{+3}$ et $-COO^{-}$. En phase gazeuse, c'est par contre la forme neutre du peptide qui est préférée [16].

1.1.3 Structures des protéines

On distingue quatre niveaux structuraux pour une protéine, respectivement appelés structure primaire, secondaire, tertiaire et quaternaire. Brièvement, la structure primaire est la séquence en acides aminés de la protéine, la structure secondaire est un premier niveau de structuration locale, la structure tertiaire est le deuxième niveau de structuration, il opère à plus longue distance (c'est à dire entre acides aminés plus éloignés le long de la séquence) et la structure quaternaire est l'assemblage de plusieurs chaînes protéiques entre elle. Toutes les protéines n'ont pas de structure quaternaire et beaucoup fonctionnent en tant que monomère. Par contre, elles ont toutes une structure primaire, secondaire et tertiaire. Dans la suite, j'illustrerai ces différents niveaux avec la toxine du choléra comme exemple et plus précisément le pentamère de la sous unité B de la toxine ($CtxB_5$). Les coordonnées de la structure atomique de cette protéine peuvent être obtenues librement sur le site RCSB de la protein databank (PDB) sous le nom de code PDB « 1EEI ».

1.1.3.1 Structure primaire

La structure primaire est l'enchaînement linéaire des acides aminés dans un peptide (Figure 1.5). Cette structure est aussi appelée séquence car il ne s'agit pas seulement d'une composition en acides aminés mais bien d'un ordre précis dans lequel les acides aminés sont covalemment liés les uns aux autres. En effet, pour une composition en acides aminés donnée, il existe plusieurs séquences alternatives possibles, mais toutes ne mènent pas à une protéine fonctionnelle. Par exemple, une partie de protéine composée d'une alanine, d'une valine et d'une leucine, peut avoir comme séquences alternatives, AVL ou ALV ou LVA, etc.

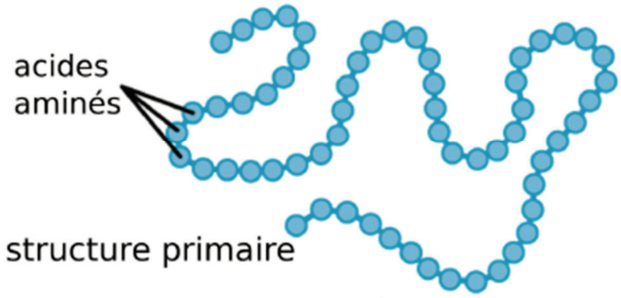


Figure 1.5 : Structure primaire [17].

La structure native d'une protéine est définie par sa séquence. Les liens entre une séquence d'acide aminés et une structure spatiale font le sujet de nombreuses recherches.

1.1.3.2 Structure secondaire

La structure secondaire résulte d'un repliement local de la protéine créé par des interactions stériques et électrostatiques et stabilisée par des liaisons hydrogène entre atomes de résidus proche le long de la séquence. Cette structure a été découverte en 1951 par Pauling, Corey et Branson [18, 19] qui ont déterminé plusieurs motifs structuraux caractéristiques. On distingue trois angles de torsion principaux dans un résidu (Figure 1.6) :

- Φ, angle C-N-C_α-C;
- Ψ, angle N-C_α-C-N.
- ω, angle C_α-C-N-C_α.

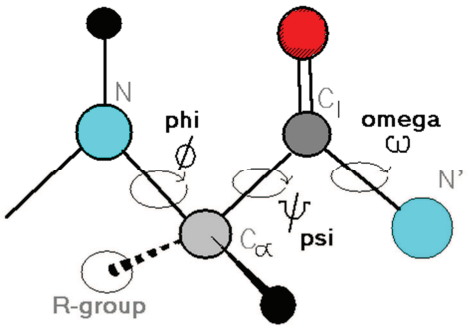


Figure 1.6 : Angles de torsion dans un peptide [20].

En raison de la double liaison partielle de la liaison peptidique, la majorité des liaisons peptidiques dans des structures de protéines sont soumises à une conformation plane [21]. L'angle ω (C_α-C-N-C_α) est bloqué sur une valeur proche de 180°, ce qui maintient la liaison peptidique dans cette conformation. L'angle χ définit la rotation de la chaîne latérale du résidu (Fig. 1.7). Il existe deux conformations possibles le long de la liaison peptidique, *trans* ou *cis* en fonction de la position relative des chaînes latérales et l'hydrogène. Cependant, la conformation *trans* est majoritairement observée à cause de l'encombrement stérique peu favorable à la conformation *cis*. Il existe une exception importante, le cas de la proline. Le résidu proline du fait de sa chaîne latérale qui se raccroche à l'atome d'azote par une liaison covalente, peut lui exister sous les deux conformations [22]. De ce fait, la proline peut influer de façon notoire sur la position du domaine qui la suit selon qu'elle adopte l'une ou l'autre des conformations (Figure 1.7). L'isomérisation *cis-trans* de la proline est souvent une étape

limitante du repliement [23-25]. La sous unité B de la toxine du choléra contient deux résidus de proline (P53 et P93) de conformation *cis* pour la proline 93 *trans* pour la proline 53 (Figure 1.7).

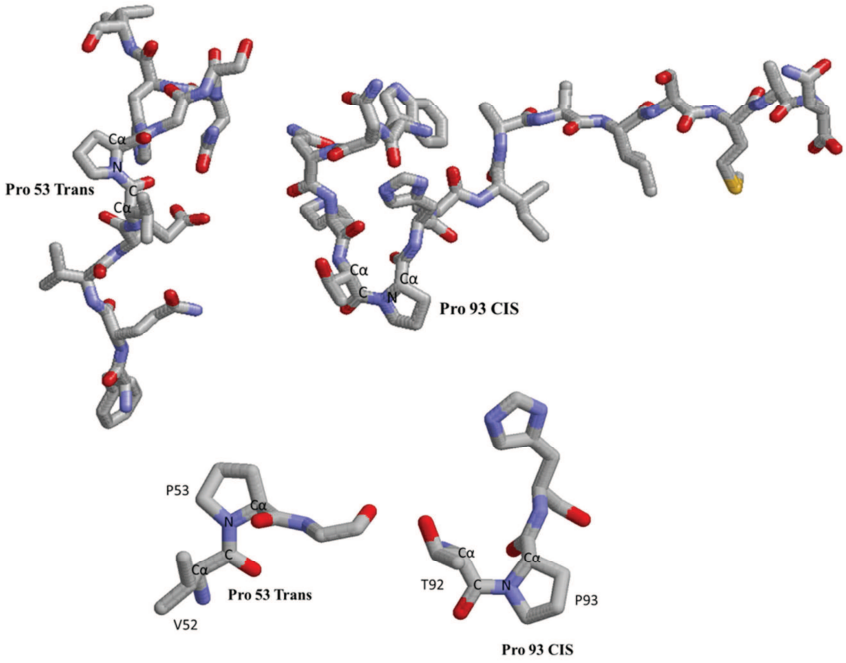


Figure 1.7 : Représentations des conformations *cis* et *trans* de la proline 93 et la proline 53 dans le pentamère B de la toxine du choléra (1EEI).

Les changements de conformations locales associées aux isomérisations *cis-trans* sont significativement différentes de celles observées dans les structures généralement similaires [26, 27].

Diagramme de Ramachandran

La chaîne polypeptidique est contrainte par l'encombrement des atomes qui composent les chaînes latérales des acides aminés ; principalement par le carbone β qui est le carbone de la chaîne latérale directement lié au carbone α (sauf pour l'acide aminé glycine dont la chaîne latérale est uniquement constituée d'un hydrogène). Pour cela, les paires d'angles Φ et Ψ ne peuvent pas prendre toutes les valeurs possibles. Le physicien indien G.N. Ramachandran a décrit la conformation locale de la chaîne polypeptidique des protéines par ces angles Φ et Ψ, permettant une représentation sous forme de graphe à deux dimensions maintenant appelé carte de Ramachandran (Figure 1.8). La figure 1.8 montre bien qu'il existe des zones « non-privilégiées» pour le repliement de la chaîne polypeptidique. Les régions favorables de cette carte contiennent des zones plus petites qui correspondent aux valeurs des

angles Φ et Ψ que l'on trouve dans les éléments de structure secondaire comme les hélices α et les feuillets β.

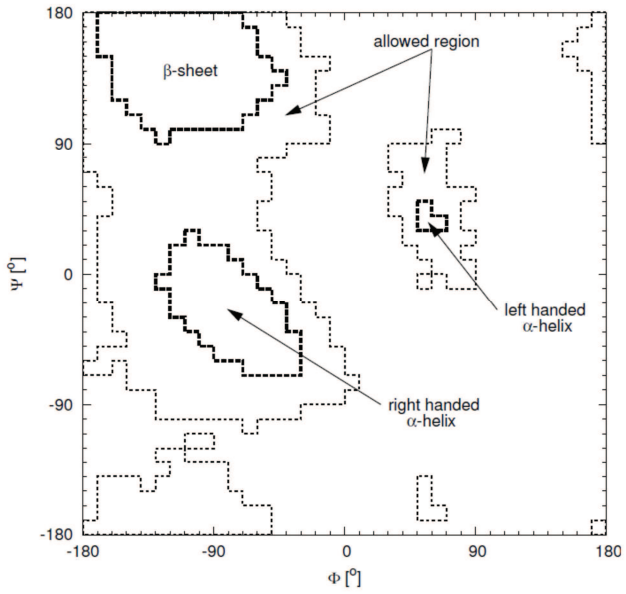


Figure 1.8 : Carte de Ramachandran avec les différentes structures secondaires associées.

Il apparaît que la chaîne polypeptidique se replie en formant ses motifs périodiques appelés éléments de structure secondaire dans quasiment toutes les protéines. Sachant que les protéines majoritairement étudiées sont globulaires. En effet les protéines désordonnées sont pauvres en structures secondaires. Ces motifs sont des régions continues, de quelques résidus à plusieurs dizaines de résidus; ils sont séparés par des régions qui possèdent une structure apériodique appelées boucles. En général, environ la moitié des résidus d'une protéine forment des éléments de structure secondaire. Il existe principalement deux types d'éléments de structure secondaire : l'hélice α (souvent désignée par la lettre H pour hélice) et le feuillet β (souvent désigné par la lettre E) représentés sur la figure 1.9.

- ***L'hélice alpha*** : Dans la structure dite en hélice alpha (α), la chaîne d'acides aminés prend la forme hélicoïdale. L'hélice alpha est stabilisée par des ponts hydrogènes établis entre l'hydrogène d'un groupement aminé -NH et l'oxygène d'un groupement carboxylique -C=O et situé quatre résidus plus loin à une distance de 5,4 Å.
- ***Le feuillet bêta*** : Dans un feuillet bêta (β), il se forme des liaisons hydrogènes entre brins β qui constituent des segments de la chaîne disposés parallèlement ou antiparallèlement les uns par rapport aux autres. L'ensemble forme comme une membrane plissée parallèle ou antiparallèle [28] (Figure 1.9).

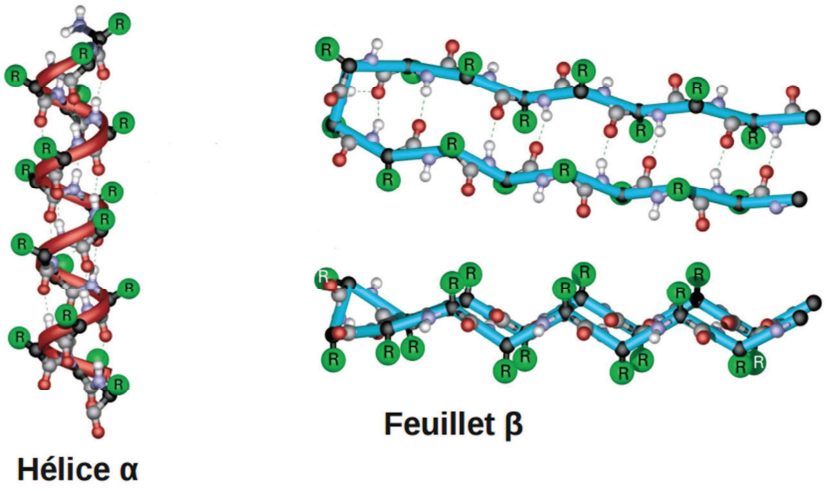


Figure 1.9 : Représentation des structures secondaires sans les chaînes latérales des acides aminés. Hélice α et Feuillet β à deux brins.

Une hélice alpha résulte de la succession d'angle Φ et ψ de valeurs approximative -57° et ~-47°, respectivement. On compte 3,6 résidus dans un tour d'hélice pour une longueur de 5,4 Å [29].

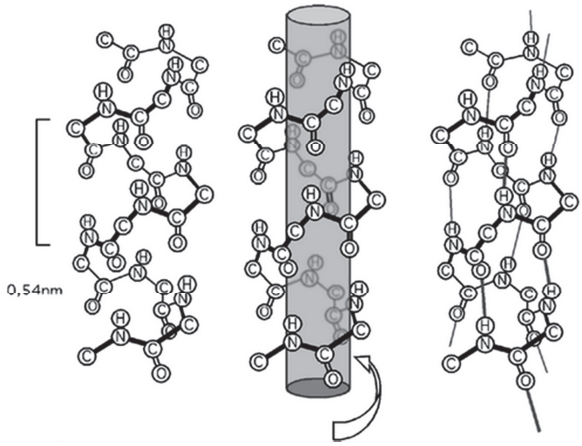


Figure 1.10 : Un tour d'hélice avec la liaison hydrogène entre les résidus i et i + 4.

La stabilité des hélices est due à la liaison hydrogène entre l'atome d'oxygène d'un résidu i et l'hydrogène du groupe NH d'un résidu i+4 (Figure 1.10). La longueur d'une telle liaison avoisine à 2,86 Å de l'atome d'oxygène à l'atome d'azote [29].

La structure en feuillet β [30] (Figure 1.11) est constituée de chaînes polypeptidiques très étirées (par exemple, Φ = -139° et = +135° pour le feuillet antiparallèle) en comparaison à la structure en hélice. En moyenne, 20 % des résidus dans les protéines sont en feuillet [31].

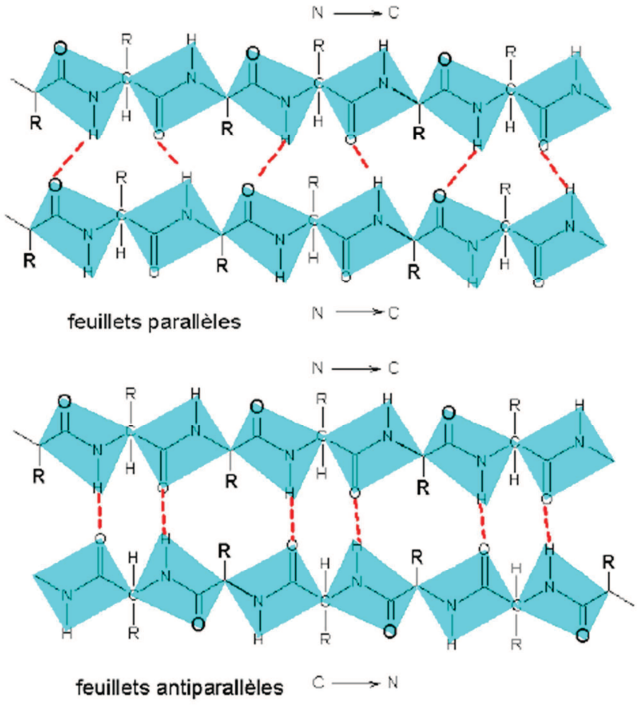


Figure 1.11 : Feuillets β parallèles et antiparallèles [32].

Les différents éléments de structures secondaires s'agencent entre eux à travers des interactions entre atomes d'acides aminés plus ou moins éloignées le long de la chaîne pour former un repliement et la forme 3D de la protéine. Les régions, dites « boucles », entre éléments de structures secondaires types hélice α et feuillet β participent à l'acquisition de la forme native d'une protéine en influant sur leur positionnement relatif. Ces boucles sont bien plus que ça, puisqu'elles ont aussi un rôle dans la dynamique et la fonction des protéines [33].

1.1.3.3 Structure tertiaire

Les protéines sont des éléments fonctionnels essentiels des organismes vivants. Les protéines accomplissent leurs fonctions biologiques en produisant des structures tridimensionnelles stables adaptées. La fonction d'une protéine est donc intrinsèquement liée à sa structure. Les atomes qui composent la protéine interagissent entre eux, et avec le solvant qui entoure la macromolécule de manière à contraindre la chaîne à se replier et à adopter une structure tridimensionnelle. Cette structure appelée « état natif » est la conformation qui va lui permettre de réaliser la fonction biologique pour laquelle elle est produite. Récemment la notion de conformation unique pour une fonction donnée a été revisitée vers la notion d'ensemble conformationnel permettant de pourvoir à une fonction [34]. Cette nouvelle

définition rend compte de la dynamique d'une protéine et de la plasticité structurale naturelle nécessaire à sa fonction.

Après une succession de repliements locaux, la protéine adopte un repliement global qui lui donne sa forme finale et son activité biologique. De façon simplifiée, la structure d'une protéine apparait hiérarchisée avec des interactions locales entre résidus voisins le long de la chaîne qui mènent aux structures secondaires et des interactions entre résidus au-delà des voisins linéaires directs, qui agissent sur la formation des structures tridimensionnelles. La séquence des différentes étapes n'est pas toujours la même pour toutes les protéines et la structuration ne suit pas la chaîne dans un sens donnée. Il s'agit plutôt d'un ensemble de réactions concomitantes, évoluant de façon très dynamique au cours du temps. Cependant la structure s'acquiert (ou se disloque) par des interactions locales qui se propagent globalement au sein de la protéine.

Le repliement conduit à l'enfouissement des acides aminés hydrophobes de la protéine dans le cas des protéines solubles dans l'eau et inversement dans le cas des protéines membranaires. La structure tertiaire d'une protéine peut contenir plusieurs motifs structuraux secondaires comme des hélices ou des feuillets (Figure 1.12).

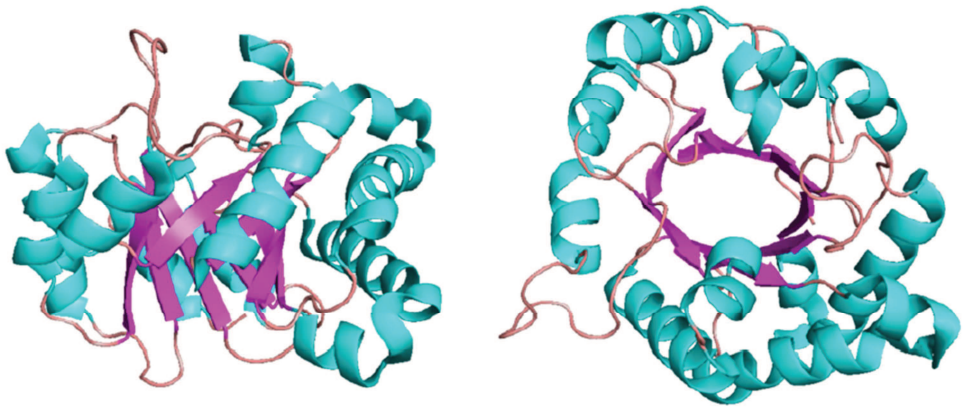


Figure 1.12 : Structure tertiaire de la sous unité B de la toxine du choléra. Les hélices sont vertes et les feuillets roses.

La structure tertiaire est stabilisée par des interactions non covalentes et par des interactions covalentes entre les atomes de soufre de résidus de cystéine lorsqu'ils sont présents, suffisamment proche et dans des conditions redox favorables. La connaissance de la structure tertiaire d'une protéine permet de connaître la position de ses atomes dans l'espace. La fonction d'une protéine est intrinsèquement liée à sa structure tertiaire. La détermination

de la structure tertiaire est donc d'une importance capitale. Les approches réseaux ont été appliquées pour comprendre la dynamique des structures tertiaires [35, 36]. Vendruscolo et al. [35] ont construit des graphiques correspondant aux structures de protéines. Ils ont montré un ensemble limité de sommets dans un graphe « petit monde » avec une grande connectivité. Ces sommets correspondent aux résidus jouant le rôle de « hubs » (acides aminés ont plusieurs liaison avec les acides aminés de la chaine adjacente) dans le réseau d'interactions.

1.1.3.4 Structure quaternaire

La structure quaternaire est le nombre de monomères associés entre eux dans un oligomère (Figure 1.13). Les zones de contact entre les monomères sont appelées des interfaces. Les interfaces sont donc faites d'interactions intermoléculaires (interactions entre atomes de deux chaînes différentes) contrairement aux structures secondaires et tertiaires qui font appel à des interactions intramoléculaires (interactions entre atomes d'une même chaîne). Aussi, il existe des oligomères covalents ou les chaînes sont associées entre elles par des ponts disulfures.

La plupart des protéines sont composées de plus d'une chaîne polypeptidique. En outre, beaucoup de protéines interagissent avec d'autres pour former transitoirement des complexes binaires impliqués dans différents processus cellulaires. En effet, la fonction biologique d'une protéine peut être considérée comme définie par le cadre de ses interactions dans la cellule, et les interactions inappropriées peuvent conduire à des maladies [37].

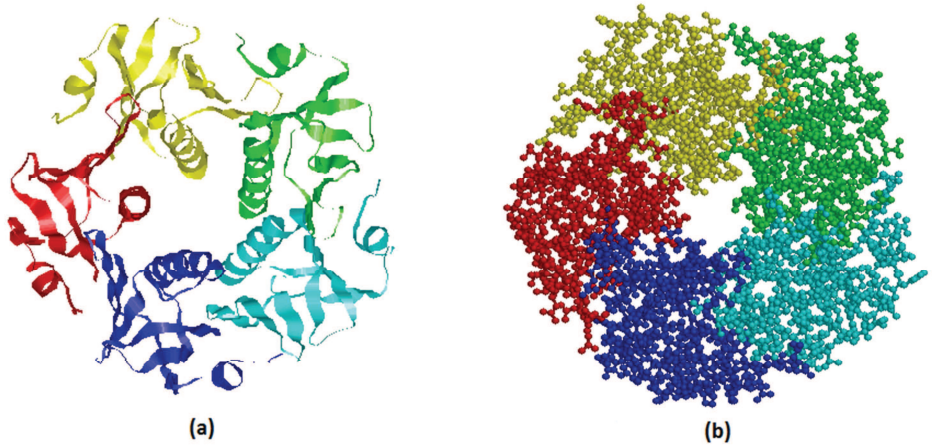


Figure 1.13 : Structure quaternaire du pentamère de la sous-unité B de la toxine du choléra (PDB 1EEI). Chaque chaîne est indiquée par une couleur différente. (a) : chaque chaîne est représentée par des rubans qui permettent de voir seulement les atomes du squelette. (b) : la protéine est représentée par des sphères et des bâtons et qui constituent les atomes (chaines principale et latérale) et les liens covalents entre eux.

1.1.4 Constituants de base des protéines : les acides aminés

1.1.4.1 Définition

Les acides aminés constituent les briques de base des protéines. Les vingt acides aminés partagent une même structure de base (Figure 1.14) comprenant une fonction amine (NH_2), une fonction acide carboxylique (COOH), une chaîne latérale (notée R sur la figure 1.14) et un atome d'hydrogène, le tout articulé autour d'un atome de carbone asymétrique (C). La nature du groupement latéral R différencie les acides aminés (Annexe 1) et est responsable de leurs propriétés (acidité, basicité, aromaticité, structure. . .).

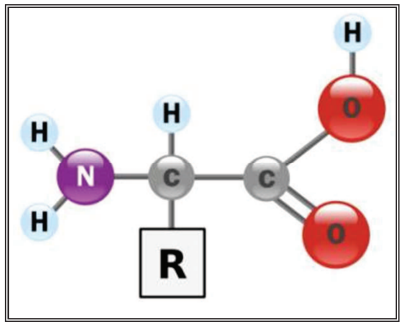


Figure 1.14: Représentation d'un acide aminé. Tous les acides aminés sont composés d'un même squelette et distingués par des radicaux R différents (chaîne latérale) qui leur confèrent leurs propriétés chimiques et topologiques.

1.1.4.2 Classification des acides aminés

Les acides aminés sont regroupés par familles en fonction des propriétés chimiques du radical (Annexe 1). Les chaînes latérales des acides aminés permettent de classer ces derniers en différents groupes partageant certaines caractéristiques. Certains sont polaires, et donc hydrophiles, d'autres non-polaires et donc hydrophobes. Certains sont chargés et ont un état de protonation dépendant du pH du milieu. Certains contiennent un cycle aromatique et présentent des propriétés spectroscopiques. Certains contiennent du soufre.

Plusieurs types de classement sont possibles puisque certains acides aminés peuvent combiner plusieurs caractéristiques. Le diagramme de Venn est utilisé [38] (Figure 1.15) afin de représenter l'ensemble de leurs propriétés. On peut donc rapidement comprendre que les acides aminés sont des objets chimiques et géométriques d'une grande versatilité et dont les réactions chimiques varieront énormément en fonction de leur environnement. Il est important de se rendre compte que tous les acides aminés sont composés d'atomes d'hydrogène, d'oxygène, de carbone, d'azote et certains ont aussi de soufre. Ce qui confère aux acides aminés leurs caractéristiques distinctes est l'agencement de ces atomes entre eux.

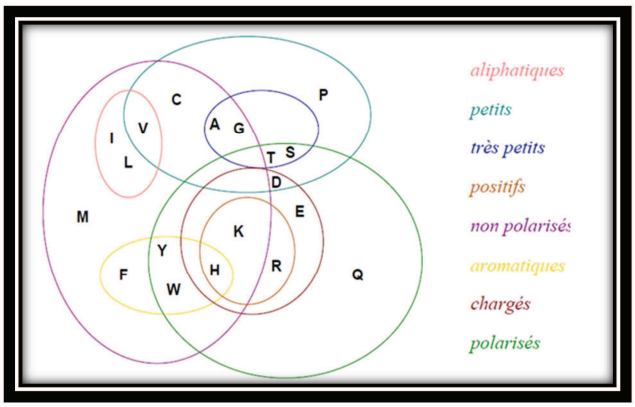


Figure 1.15 : Diagramme de Venn des propriétés des acides aminés.
Les différentes catégories sont faites en fonction de la nature de la chaîne latérale R. Trois grandes catégories se chevauchent, les hydrophobes, les polaires et les petits. Il existe également plusieurs sous-catégories, les aliphatiques, les aromatiques, les neutres, les chargés positivement, les charges négativement, les minuscules et la proline. Certaines acides aminés, telle la glutamine, n'appartiennent qu'a une seule catégorie, d'autres telle la thréonine, qui fait partie des trois catégories majeures, partagent plusieurs appartenances.

1.1.4.3 Propriétés des acides aminés

Les acides aminés ont des propriétés physico-chimiques très diverses. Concernant les propriétés physiques, je peux citer : la solubilité, l'optique, l'absorption lumineuse dans l'ultraviolet… Les propriétés chimiques des acides aminés sont nombreuses grâce à la présence au sein d'une même molécule de groupements carboxyle, aminé et de groupements variés sur la chaîne latérale (carboxyle, amine, alcool, amide, phénol, thiol etc.). La fonction acide carboxylique (-COOH) présente des réactions telles que : formation de sel, réduction, estérification, décarboxylation. Les propriétés chimiques dues à la fonction amine primaire – NH2 sont : substitution par un radical R, désamination, transamination…

1.1.4.4 Différents types de liaisons entre les acides aminés

Les vingt acides aminés naturels partagent quatre atomes formant le squelette de la protéine et se distinguent par les atomes de la chaîne latérale. Ces atomes peuvent faire différents types de liaisons chimiques (Figure 1.16). Tout d'abord, les acides aminés sont liés les uns aux autres par une liaison covalente comportant deux atomes du squelette, (C et N) et appelé la liaison peptidique comme décrit auparavant. Les liaisons covalentes sont ainsi utilisées pour fabriquer une chaîne d'acides aminés disposés dans un ordre spécifique, la séquence primaire de la protéine, comme mentionné auparavant. Les atomes des chaînes latérales sont aussi liés entre eux au sein d'un seul acide aminé par des liaisons covalentes. Il peut y avoir des liaisons disulfures intramoléculaires (entre deux résidus de cystéine d'une chaîne) ou des liaisons disulfures intermoléculaires (entre deux cystéines, chacune située sur

une chaîne distincte), ce dernier faisant un oligomère covalent. Les liaisons covalentes sont des liaisons fortes car il faut une grande quantité d'énergie pour les casser (50-110 kcal / mol). Dans les organismes vivants, une enzyme (protéase) est nécessaire pour couper une liaison covalente.

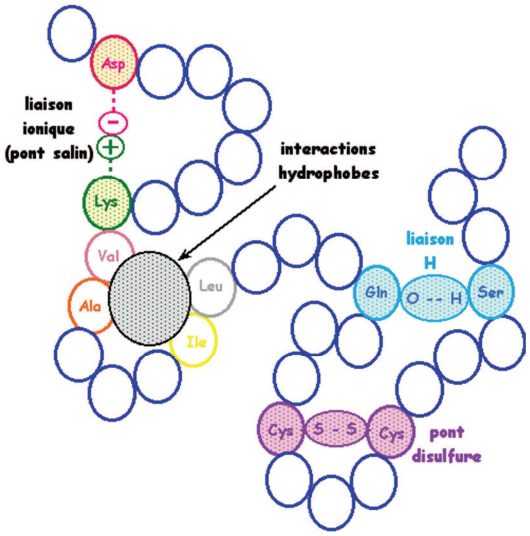


Figure 1.16: Les différents types d'interaction entre les acides aminés.

Les liaisons faibles, qui nécessitent peu d'énergie pour se casser (1-7 kcal / mol), sont impliquées dans la formation des structures secondaires, tertiaires et quaternaires des protéines lors du repliement et de l'association des chaînes. Ce sont des liaisons hydrogènes, liaisons hydrophobes, des liaisons électrostatiques (entre les charges), polaires (entre dipôles) et interactions de Van Der Waals. Dans des conditions physiologiques, les liaisons faibles se forment et se brisent continuellement. Les structures secondaires des protéines, des α-hélices et les feuilles β (interactions de β-brin intramoléculaires) sont stabilisées par des liaisons hydrogène entre les atomes du squelette des acides aminés. Ces liaisons se font entre les acides aminés composant une hélice α mais entre acides aminés de deux brins β différents.

1.1.5 L'interface protéique

Les interactions intermoléculaires entre deux chaînes associées forment une zone de contact qui prend le nom d'« interface ». L'étude de l'interface a pour objectif d'identifier les résidus qui interagissent avec une autre molécule [39, 40].

Chaque chaîne offre un domaine qui reconnaît un autre domaine, ou le même domaine, sur une autre chaîne. L'association est basée sur les complémentarités chimiques et géométriques des deux domaines. Ces complémentarités sont construites sur la disposition spatiale des interactions intermoléculaires des acides aminés qui suivent des motifs permettant

la reconnaissance des deux domaines. Un motif très étudié et bien compris est la super hélice (‘coiled-coil’) qui est composé de répétitions de sept acides aminés consécutifs appelés ‘heptead’ et décrites par les lettres *abcdefg* dans lesquels les interactions entre résidus *a* et *d* de chacune des hélices servent d’ancrage pour former l’interface [41]. Il existe de nombreux autres motifs de reconnaissance entre domaines permettant la formation d’une interface, ces motifs sont répertoriés dans différentes bases de données [42-44]. Pourtant, les règles qui nous permettraient de prédire les séquences des interfaces restent encore insaisissables. Les interfaces de protéines ont été largement étudiées [39, 40]. Mais parce que les interfaces sont grandes et de nature assez planaires, elles manquent des contraintes géométriques réalisées par un nombre limité de séquences. D’où l’absence de motifs triviaux ou de profils identifiables sur la séquence au contraire des sites actifs enzymatiques. Ce problème est un de sujets d’étude de ma thèse (chapitre 7 et 8) [45, 46].

Cependant il est nécessaire de bien comprendre ces motifs de reconnaissance pour pouvoir construire des inhibiteurs d’assemblage efficace, essentiels au développement thérapeutique contre les maladies liées à des mauvais repliements ou des mauvais assemblages comme la maladie d’Alzheimer. La compréhension détaillée de ces motifs est aussi importante simplement pour permettre à partir de la séquence d’une protéine de déterminer si elle existe sous forme oligomérique ou non, et le cas échéant sa structure quaternaire. En effet, il reste toujours difficile de déterminer expérimentalement de façon certaine la présence d’oligomères. La théorie des graphes est un outil idéal pour étudier les résidus impliqués dans les interactions moléculaires, les résidus impliqués dans les interactions intermoléculaires et leurs communications [47].

1.1.5.1 Formation des interfaces protéiques

Les protéines sont composées de l’association de petits fragments responsables à la formation des noyaux de nucléation définies par certains auteurs sous le terme de « hot spots ». Cette découverte a conduit à un grand développement d’outil de calcul visant à identifier les hots spots (Les acides aminés essentiels pour la formation de l’interface sont maintenant connus familièrement comme hot spot).

a. Composition des « Hot spots »

Les « hot spots » sont des acides aminés situés dans l’interface. Bogan et Thorn ont étudié l’enrichissement de chaque type d’acide aminés dans les « hot spots » par rapport au reste de la surface sur une base de données[50]. Dans cette étude, seuls les acides aminés dont la surface accessible au solvant est supérieure à 10 Å² sont considérés comme quoi hot spots. La comparaison de la fréquence de chacun des acides aminés d’un ensemble de 2325 résidus

avec la fréquence des acides aminés présents dans les « hot spots » de ces protéines a été étudiée. Les résultats ont définis les « hot spots » de leur jeu de données en étudiant par « alanine scanning » les différences d'énergies d'interaction, et en ne gardant que les résidus dont la contribution était supérieure à 2 kcal. L'hypothèse était que les acides aminés favorisés dans les « hot spots » sont ceux qui sont capables de réaliser plusieurs types d'interactions favorables formant l'interface. Des résultats différents à ceux présentés précédemment [42] ont montré que trois acides aminés, le tryptophane, la méthionine et la phénylalanine sont beaucoup plus représentés à l'interface. Cela a été analysé en comparant la proportion de chaque type d'acides aminés conservés à la surface des protéines avec celle des types d'acides aminés conservés dans la zone d'interaction avec le partenaire protéique. La présence de ces résidus peut être utilisée pour prédire les zones d'interactions avec un partenaire protéique.

b. Détection des « hot spots »

Expérimentalement et classiquement, la recherche de « hot spots » se fait par l'évaluation du changement de l'énergie libre (le travail qu'on peut extraire de la réaction) du complexe protéine-protéine lorsque l'on mute les acides aminés de l'interface en alanine. C'est par cette méthode que les « hot spots » ont étés définis pour la première fois [51]. Deux bases de données sont utilisées pour réaliser des mutations en alanine sur des complexes protéiques : BID (1300 mutations disponibles pour 170 complexes protéiques) [52] et ASEdb (2915 mutations pour 91 complexes protéiques) [53].

Pour déterminer la présence de « hot spots », une approche informatique génère *in silico* les mutations des acides aminés de l'interface en alanines en utilisant la structure tridimensionnelle des protéines, suivies d'une évaluation des variations de l'énergie libre théorique du complexe. Cette approche basée une méthode énergétique en utilisant l' « alanine scanning » a été intégrée dans le logiciel FOLDEF [54] et a été mise à disposition de la communauté scientifique par le biais du serveur Fold-X.

D'autres études sont réalisées pour prédire les « hot spots » par alanine scanning [55-57]. Des logiciels (PNAS, DrugScorePPI, …) mutent *in silico* dans un premier temps les acides aminés de l'interface en alanine et calculent ensuite la différence d'énergie libre avec l'état natif, grâce à une équation prenant en compte la somme de la contribution des énergies libres. Parmi les serveurs utilisant ces algorithmes, les plus connus sont Robetta [58]et Fold-X [59].

En plus des algorithmes énergétiques, il existe les algorithmes basés sur des paramètres structuraux, et des algorithmes basés sur l'évolution. Parmi les algorithmes utilisant des

paramètres structuraux, KFC2 [60] permet de prédire les « hot spots » grâce à 47 paramètres (dont l'ASA) et un algorithme de machine à support de vecteur. Les algorithmes utilisant l'évolution ne nécessitent pas d'information structurale sur la surface de la protéine, mais uniquement la structure primaire. Le serveur ISIS utilise la séquence d'acides aminés sans information sur la structure tertiaire de la protéine, ni sur le partenaire protéique [61].

1.1.5.2 Descripteurs des interfaces protéiques

Les chaînes latérales ont des propriétés moyennes chimiques similaires aux acides aminés des chaînes entières. La géométrie des chaînes latérales d'acides aminés est un paramètre clé pour comprendre les paires correspondant. Pour cela, la notion de réseau a été explorée.

La conception de calcul de nouvelles interactions est une application importante qui a rencontré un succès remarquable ces dernières années [62]. La modélisation thermodynamique et cinétique des interactions protéines présentent un défi. La relation entre la structure et l'affinité a été étudié. L'affinité repose sur des interactions non covalentes qui maintiennent un complexe protéique et qui sont les mêmes que celles impliquées dans le repliement des protéines [63].

Les interactions antigène-anticorps sont l'équivalent moléculaire d'une première rencontre avec une constante de liaison de l'ordre de 10^{-9} mol^{-1}. Pour certains complexes dimériques, cette constante peut avoir des valeurs de l'ordre de 10^{-16} mol^{-1} et on est donc en présence d'interactions aussi fortes qu'à l'intérieur d'un monomère, il faut dénaturer le dimère pour casser la liaison. Les complexes enzyme-inhibiteur montrent un caractère intermédiaire avec des constantes de liaison de l'ordre de 10^{-7} mol^{-1} à 10^{-13} mol^{-1}[64].

Sur la base des caractéristiques de séquence, Ofran et Rost [65] ont analysés les structures des interfaces protéiques en divisant celle-ci en 6 types distincts : interaction intramoléculaires entre domaines, intermoléculaires pour les complexes homomériques permanents, homomériques transitoires, hétéromériques permanents et hétéromériques transitoires. Chaque type d'interface correspond à une association fonctionnelle ou structurelle entre les résidus différents. Les différences entre les six types étaient si importante que, en utilisant la composition de l'acide aminé seul, la prédiction statistique d'un set de donnée à laquelle des six types d'interfaces de 1000 résidus appartient à 63-100% de précision. L'originalité de ce travail réside dans la mise au point d'une méthode permettant de classer efficacement de grandes bases de données en deux types d'interfaces (transitoires et

permanentes), types connus pour être impliqués dans la diversité des interactions protéine-protéine [66, 67].

Afin de bien déterminer si un résidu est en surface protéique, il est important de définir deux notions: la surface accessible au solvant (‘Accessible Surface Area’, notée ASA) et la surface relative accessible (notée relative ASA). Le concept de surface accessible au solvant est un concept largement employé par les scientifiques qui étudient les protéines au niveau atomique. La description de l’ASA remonte à 1971 avec les travaux et la création de l’algorithme de Lee et Richards [68].

Les analyses des séquences n’aboutissent pas à proposer des motifs spécifiques des interfaces du fait de la planéité et du manque de contrainte géométrique. Donc, l’essor des structures 3D a ouvert la porte à des analyses de la structure 3D. On a vu que les mêmes géométries/structures existent aussi dans les interfaces. L’analyse systématique de toutes les interfaces a été réalisée en utilisant des approches informatiques [70]. Plusieurs algorithmes ont été élaborés afin de classifier la structure des interfaces protéine-protéine, permettant la création de bases de données comme par exemple, Structural Classification Of Protein–Protein Interfaces SCOPPI [46], Structural Characterization Of Water, Ligands and Proteins SCOWLP [71], Protein-protein Interaction Prediction by Structural Matching PRISM [72]. SCOPPI est une base de données complète qui classe et annote les interactions de domaine provenant de toutes les structures de protéines connues. SCOWLP permet de caractériser et de visualiser la structure 3D des interfaces protéiques. PRISM est un algorithme qui cherche les interactions binaires possibles entre les protéines à travers la similitude de structure et la conservation évolutive d’interfaces connues.

1.1.6 Relation : séquence–structure–fonction

Comme mentionné précédemment, la fonction des protéines découle de leur structure tridimensionnelle, et l’activité biologique ne peut s’exercer qu’à partir de l’état replié de la protéine. L’état replié dépend de la fonction de la protéine, il signifie que la protéine n’est plus dans un état juste après synthèse. La connaissance de la structure des protéines est donc essentielle à la compréhension de leur mécanisme de fonctionnement et facilite la perception quant à leur implication dans certains processus biologiques fondamentaux. Le repliement d’une protéine et l’acquisition de sa structure font l’objet d’un contrôle cellulaire rigoureux afin de garantir qu’une protéine produite soit fonctionnelle dans le plus souvent des cas. Ce contrôle s’établit à deux niveaux, dans la séquence elle-même (contrôle interne) et dans l’environnement (contrôle externe) par le biais d’autres protéines aidant ou contrôlant la qualité du repliement. Une séquence résulte de la pression de l’évolution, elle est de ce point

de vue un objet bien contrôlé. J'aborde les mécanismes impliqués dans le contrôle interne dans les chapitres 8, 10 et 11 par l'analyse de l'impact structural de mutations *in silico*.

Le contrôle interne signifie aussi que la protéine est capable de supporter des changements sans nécessairement d'impact négatif. Par exemple, il existe de nombreuses mutations qui sont tolérées puisqu'elles n'induisent pas de perte de fonction. Ces changements de séquence n'entraînent pas de changements structuraux tels que la protéine ne puisse plus fonctionner ou plus se replier. Ces résultats suggèrent que seuls certains acides aminés dans une séquence contrôlent le repliement d'une protéine. Ceci est en parfaite adéquation avec le fait que le repliement n'est pas une exploration complète de l'ensemble de toutes les conformations possibles associées à une séquence données mais une exploration guidée. L'identification des acides aminés régulant spécifiquement le repliement et l'assemblage s'il y a lieu est néanmoins difficile. C'est un problème que j'étudierai en me reposant sur des approches informatiques, relevant en particulier de la théorie des graphes. La forme globale d'une protéine est définie par un ensemble d'interactions chimiques entre atomes des acides aminés (information locale) et les changements de formes inhérent à sa fonction (ex. repliement, assemblage ou allostérie) résultent de modifications de cet ensemble. On a, donc, à faire à un système complexe, où plusieurs éléments distincts (information locale) interagissent entre eux (forme globale) de façon dynamique (changement de forme). La modélisation par un graphe est extrêmement pertinente puisqu'elle permet de considérer dans un modèle unique, l'information locale fourni par les éléments de bases (acides aminés/atomes), l'information globale constituant le réseau (structure 3D/4D de la protéine) et la dynamique du système (changements de forme).

1.2 Le mécanisme d'assemblage protéique

1.2.1 Le mécanisme d'assemblage par des approches expérimentales

L'assemblage de protéines ou l'oligomérisation nécessite des réactions de repliement et d'association. Ainsi, pour avoir une image complète du mécanisme d'assemblage, en plus d'étudier les interactions intermoléculaires d'acides aminés, il est nécessaire d'étudier les interactions intramoléculaires d'acides aminés et d'appréhender comment ces deux types d'interactions sont coordonnés.

Jusqu'à présent, deux façons de former un dimère ont été observées. Expérimentalement, le modèle dit « induced-fit » ou « lock and key » (clé-serrure) et le modèle dit de « Fly-casting » [2, 71, 72]. La première, le modèle des trois états (deux étapes) où les monomères dépliés U

(état 1) se replient (état 2), qui associe en dimères D (état 3) (Figure 1.17). La deuxième : l'itinéraire alternatif est à travers le modèle à deux états ou les monomères U dépliés (état 1) s'associent en dimères (état 2). Les interactions intramoléculaires et intermoléculaires se produisent de façon séquentielle dans le modèle des trois états, mais de façon concomitante dans le modèle à deux états. L'hypothèse suivante est donc plausible, le repliement et l'association vont être liés, mais indépendants dans le modèle des trois états et concertés dans le modèle à deux états. En termes de réseaux, le modèle à trois états suggère une protéine organisé en deux sous-graphes connectés à distance, un qui régit les interactions intramoléculaires et l'autre les réactions intermoléculaires. Au contraire, le modèle des deux états suggère deux sous-graphes connexes. Je présente des éléments de piste en faveur de cette hypothèse dans le chapitre 6.

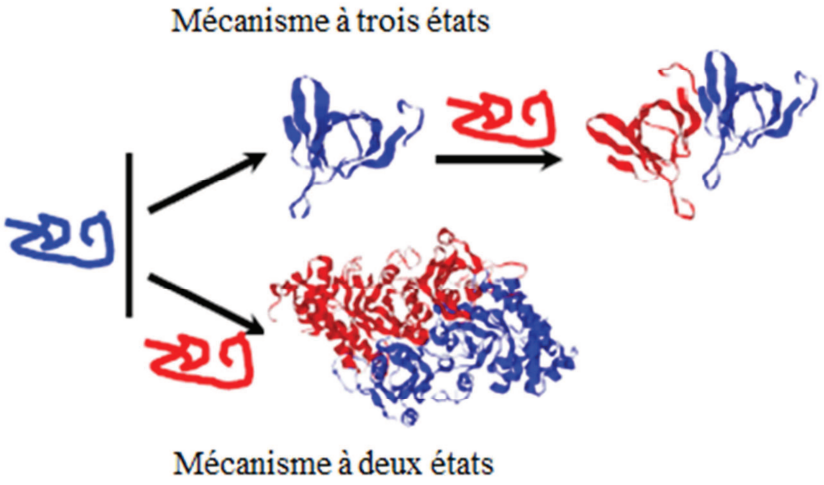


Figure 1.17 : Deux types de mécanisme pour avoir un dimère. Les chaînes protéiques sont présentées par les couleurs rouge et bleu.

Dans le modèle à trois états, l'interface dans les monomère repliés dissociés est très probablement proche de celle dans les monomères repliées et associées (dimère) puisque le repliement aura déjà positionné les acides aminés dans l'espace et contraint la géométrie des domaines pour leur rencontre au cours de la diffusion. Ainsi, il est cohérent d'utiliser la structure aux rayons X de l'oligomère de la protéine native en tant que modèle pour concevoir des inhibiteurs d'assemblage. En revanche, dans les mécanismes à deux étapes, modéliser l'interface à partir des rayons X de la structure native semble moins pertinent puisque les monomères dépliés ou partiellement dépliés s'assemblent. Il est donc fort probable que l'interface dans le dimère soit différente de celle dans le monomère et que des inhibiteurs conçus sur l'interface native ne soient pas efficaces. Ceci est une illustration pour répondre à la question : pourquoi est-il important d'anticiper le mécanisme d'assemblage ?

Il est intéressant de noter que certaines protéines de structures et séquences très proches peuvent néanmoins suivre des mécanismes d'assemblage différents [22, 23, 71].

Comme exemple, je citerai les pentamères de la toxine thermolabile B (LTB_5) et de la toxine cholérique ($CtxB_5$) qui partagent une identité de séquence de 87 % et des structures atomiques presque superposables mais s'assemblent néanmoins par deux mécanismes différents. La toxine du choléra suit un 'Fly-casting' mécanisme où les réactions de repliement et d'assemblage sont concomitantes alors que les monomères de LTB se replient quasiment sous une forme native avant de s'associer entre eux. Ces résultats renforcent l'idée que seuls certains acides aminés sont décisifs dans le repliement et l'assemblage et la comparaison des deux toxines offre un accès pour identifier ces résidus et les mécanismes qui les régissent. Ce problème sera traité en particulier dans les chapitres 4, 5 et 6 de mon manuscrit.

1.2.2 Le mécanisme d'assemblage par des approches informatiques

Le modèle à deux états a été étudié en montrant qu'une protéine dépliée à un rayon de capture plus élevé pour un site de liaison spécifique que l'état plié qui a une conformation restreinte [73]. Un mécanisme de liaison où la protéine non repliée se lie d'abord faiblement à un point d'ancrage sur le partenaire et à une distance relativement importante du site de liaison. Puis la protéine approche du site de liaison au fur et à mesure et se replie, pour enfin adopter la conformation finale et disposer correctement les résidus impliqués dans l'interaction, c'est le mécanisme Fly-casting [72]. En 2004, Wolynes met en évidence que certaines caractéristiques de la structure atomique d'un oligomère permettent d'anticiper le mécanisme d'assemblage. La taille de l'interface, son caractère hydrophobe et le rapport du nombre de contacts intermoléculaires sur le nombre de contacts intramoléculaires influent sur le mécanisme [74]. Cependant, les toxines $CtxB_5$ et LTB_5 ne peuvent se distinguer par ces paramètres suggérant que le problème est plus complexe et reste ouvert. Comme mentionné précédemment, j'ai abordé cette question dans le cadre de ma thèse et j'ai proposé des pistes alternatives en utilisant des approches réseaux (chapitre 6).

Des approches informatiques fournissent également des preuves soutenant le mécanisme clé-serrure, le mécanisme Fly-casting et une série d'entre deux mécanismes attestant d'allers/retours entre les réactions de repliement et d'association. Récemment, la dynamique moléculaire a été combinée avec l'analyse de réseau pour fournir des détails concernant la compréhension de la voie d'assemblage. Par exemple, les réseaux de transition

à gros grains peuvent être dérivés à partir de la simulation de la dynamique moléculaire afin de faire apparaître la transition entre des oligomères de différentes tailles. Maintenant, les réseaux de transition cinétiques peuvent être calculés pour les petites protéines en utilisant l'optimisation de la géométrie pour caractériser le minima des états de transition. Les réseaux peuvent être visualisés en construisant des graphiques dis-connectivité. Quand les constantes de la vitesse sont associées à des réarrangements arbitrés par chaque état de transition on peut définir un réseau de transition cinétique [75].

Il est évident que l'assemblage protéique résulte de réactions de repliement et d'association orchestrées savamment. Cela signifie que les facteurs clés pour l'assemblage de la protéine sont dû à l'équilibre entre les interactions intramoléculaires et intermoléculaires.

1.3 Les maladies liées aux mauvais repliements des protéines

Certaines maladies neuro-dégénératives sont des pathologies liées à un repliement anormal de certaines protéines ou des peptides [76]. Ces derniers sont capables de changer de conformation et de s'auto-assembler dans des morphologies type fibres. La toxicité des maladies est décrite comme fortement liée à leur processus d'auto-assemblage et à leurs interactions avec des membranes cellulaires dans lesquelles plusieurs espèces formées peuvent contribuer à la toxicité [77].

L'accumulation de protéines agrégées dans le système nerveux central est la signature moléculaire de maladies dégénératives dévastatrices comme les maladies d'Alzheimer, de Parkinson, de Huntington ou de Creutzfeldt-Jacob.

Les mécanismes responsables de la dégénérescence neuronale semblent communs à une grande majorité de ces maladies. Une caractéristique commune est l'existence dans le système nerveux d'agrégats fibrillaires de protéines « mal repliées », constituant des dépôts amyloïdes dans les cellules nerveuses ou dans les espaces extracellulaires. De manière surprenante, alors que ces molécules diffèrent par leur structure et leur fonction, les structures des agrégats sont très similaires, avec une conformation en feuillets β croisés qui leur confère un caractère d'insolubilité et probablement de toxicité. Le plus souvent la pathologie implique la formation d'un oligomère ou d'une fibre via la formation d'un feuillet β intermoléculaire. Pour comprendre pourquoi seules certaines protéines dérivaient vers des formes pathologiques via des feuillets β alors que de nombreuses protéines saines ne le font pas. J'ai étudié les propriétés de réseaux des feuillets intermoléculaires issus de protéines saines et je les ai comparées à celles connues des feuillets β intermoléculaires des protéines liées à des pathologies (Chapitre 7 et 8).

Pour chaque pathologie, une ou plusieurs protéines sont impliquées, et certaines protéines sont communes à plusieurs maladies. L'accumulation débute bien avant les premiers symptômes de la maladie.

Le phénomène est probablement lié à la synthèse d'une protéine de structure anormale du fait d'une mutation dans le gène qui gouverne sa synthèse. Dans les formes familiales, l'agrégation fibrillaire est accélérée, conduisant souvent à un début plus précoce de la maladie. Dans les formes sporadiques, en revanche, la protéine synthétisée est *a priori* normale, et les mécanismes conduisant à son agrégation ne sont qu'hypothétiques. Quel que soit le cas (forme familiale ou sporadique), un dépassement ou un défaut des « systèmes de surveillance » dans l'organisme a été incriminé (Figure 1.18) [78].

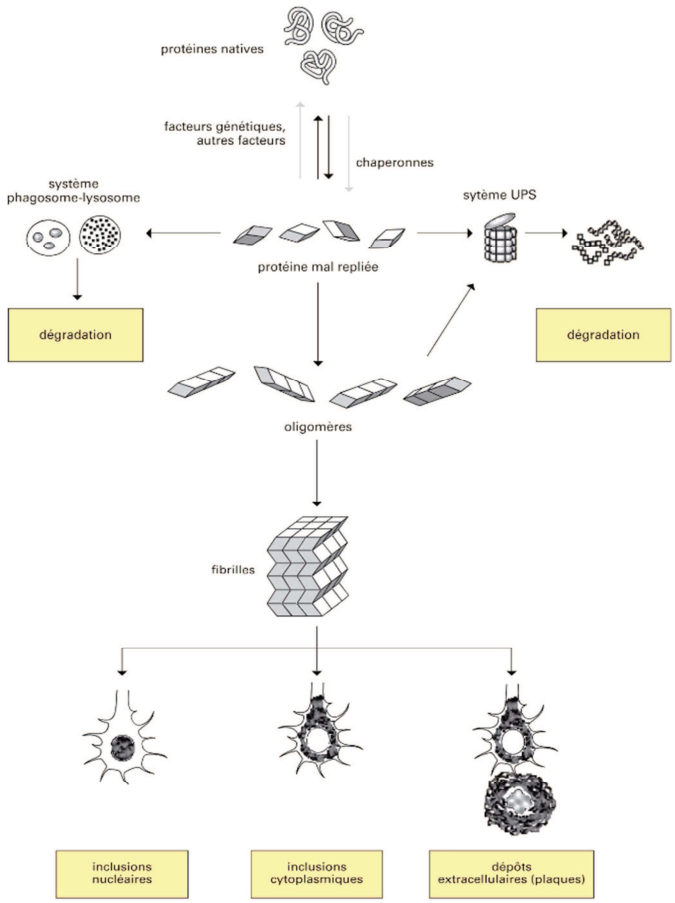


Figure 1.18 : Contrôle du bon repliement des protéines par élimination des échecs
Le système de « surveillance » (protéines chaperonnes, système ubiquitine-protéasome ou UPS, système phagosome-lysosome) est chargé d'éliminer les protéines mal repliées. Lorsque ces systèmes échouent, les protéines mal repliées s'accumulent sous formes d'oligomères ou de fibre qui vont interférer avec le bon fonctionnement de la cellule (inclusions, dépôt fibrillaire) et entraîner des maladies.

Protéines liées aux pathologies

Le tableau présente quelques exemples de protéines impliquées dans des maladies conformationelles.

Maladie	Protéine/peptide impliqué	Lésion ou effets pathologiques
Maladies neuro-dégénératives		
Maladie d'Alzheimer	Peptide amyloide béta, AB	La perte irréversible des neurones
Encéphalopathie spongiforme	Prion	Perte neuronale
Maladie de Parkinson	α-synucléine	La perte des neurones dopaminergiques dans la substance noire et la présence des inclusions intracellulaires contenant α-synucléine
Maladie de Huntington	Huntingtine-poly Q	Neurones (inclusions intranucléaires)

Tableau 1.1: Exemples et classification des maladies humaines associées à la formation de dépôts extracellulaires ou des inclusions intracellulaires [79]

Chapitre 2: Notions de réseaux

2.1 Introduction

Les réseaux sont partout. De l'Internet aux réseaux économiques, en passant par les réseaux urbains, et même les réseaux terroristes, le concept de réseau imprègne le monde contemporain. Tout système d'éléments interconnectés par des liens de n'importe quelle nature, tels que les liens entre les sites Web, les entreprises et les relations sociales entre les gens, la connectivité des générateurs, transformateurs et de postes connectés par des lignes électriques, peuvent être vus comme des réseaux. L'étude de ces réseaux multiples et variés apporte des progrès significatifs dans la compréhension de la topologie, la croissance et la dynamique de ces systèmes dits complexes [80, 81]. L'étude des réseaux devient actuellement une science à part entière [82]. Dans la suite nous nous fonderons sur la description donnée par SALAMATIAN [83] qui définit un réseau est en tant qu'« ensemble d'acteurs distincts et distribués dans l'espace qui coopèrent ensemble afin de s'échanger de l'information ».

Cette description initiale est importante pour le cas que je traite dans ma thèse, celui des réseaux dans la structure des protéines. En effet, à la différence d'autres réseaux où une information est sous forme de bits, ou un flot d'énergie, ou de produit qui transite, aucun transfert de ce type n'a lieu dans les réseaux que j'étudie. Néanmoins, en suivant la définition de réseau précédente, il existe un ensemble d'acteurs distincts et distribués dans l'espace, les acides aminés ou les atomes, qui coopèrent afin de s'échanger une information structurelle qui va aboutir à une forme particulière, par le biais du repliement (folding) et à certaines propriétés biochimiques. Cette information résulte des liens chimiques qui génèrent des contraintes géométriques et qui aboutisse à une forme 3D. Le réseau agit donc en tant qu'outil pour propager des perturbations locales à une échelle globale. Il est donc intéressant d'étudier comment cet échange d'informations de contraintes structurales impacte sur la plasticité structurelle des protéines à différentes échelles, du local vers le global : la structure protéique entière.

Les réseaux biologiques ont reçu beaucoup d'attention au cours des dernières années car ils modélisent les interactions complexes qui se produisent entre les différents composants de la cellule [84]. Grâce au développement de technologies à haut débit [85], de grands volumes de données expérimentales sur les interactions des protéines ont été mis à

disposition. Evidemment, l'élucidation du fonctionnement des réseaux d'interactions et le développement d'une approche biologique systémique nécessite des outils informatiques aidant le biologiste à raisonner, inférer et concevoir des expériences permettant d'analyser qualitativement la dynamique des réseaux d'interaction.

Les graphes sont devenus ces dernières années, notamment grâce à l'étude des réseaux sociaux, le principal outil de représentation de relations forte ou faible entre variables dans tous les champs d'application. On retrouve les graphes à divers niveaux de description des mécanismes en biologie systémique, les réseaux métaboliques (graphes entre réactions et produits de réaction) ; les réseaux protéines/protéines (graphes de proximité des séquences protéiques); les réseaux de régulation (graphes des interactions du produit d'un gène sur la transcription d'autres gènes) ; ou les réseaux de co-expression (graphes des corrélations entre niveau d'expressions de gènes). Si l'analyse des réseaux biologiques a beaucoup emprunté aux télécommunications et aux sciences sociales, domaines où la théorie des graphes est solidement ancrée, les systèmes biologiques ont généré de nouvelles problématiques dues à l'extrême complexité des mécanismes en jeu, à leur instabilité et à la nature des sources de données associées (typiquement un très grand nombre de variables pour peu d'observations). Dès lors, les statistiques se sont imposées pour gérer les incertitudes liées aux phénomènes et à l'afflux massif de données de grande dimension, rendant souvent prohibitive l'utilisation à grande échelle de modèles déterministes [86].

Deux éléments clés sous-jacents, la connectivité de nombreux réseaux du monde réel sont la propriété de «petit monde » et l'existence ou non d'échelles caractéristiques [87]. L'application de ces éléments à la sphère biologique ont révélé que les réseaux d'interactions moléculaires impliqués dans le métabolisme cellulaire, ainsi que les processus de régulation transcriptionnelle présentent des propriétés de petit monde et n'ont pas d'échelle caractéristique [81]. Des efforts ont également visé à appliquer les concepts de réseau à l'étude du repliement des protéines [35, 88] et à étudier la connectivité entre les protéines repliées. Il y a aussi beaucoup de travaux sur les protéines oligomériques. Cette thèse se positionne dans le même contexte que ces travaux. Je présente dans cette thèse une perspective alternative qui modélise une structure protéique par un réseau, et essaye de déduire des propriétés structurales des protéines grâce aux propriétés de ces réseaux. Mon travail s'intéresse aux réseaux d'acides aminés qui permettent de mieux comprendre la cellule.

A partir des connaissances sur les réseaux d'acides aminés aux interfaces peut extrapoler les partenaires protéiques dans une cellule pour un travail cellulaire. Si la fonction d'au moins un des partenaires est connue, alors on essayera d'associer sa fonction et le chemin fonctionnel dans lequel la protéine est sensée participer. Il est donc, à travers les réseaux intriqué de ces interactions que l'on peut espérer de réaliser une carte des chemins fonctionnels de la cellule, leurs inter-connectivités et leur régulation dynamique. Ainsi il faut détecter les complexes de protéines interconnectés entre eux. Ce problème se réduit en la résolution d'un problème de « clustering » sur les graphes. Le clustering consiste à regrouper des objets interconnecté par des graphes en groupes (appelés communautés) de telle sorte que les objets dans le même cluster sont plus interconnecté entre eux qu'aux objets des autres classes [89]. Ainsi des groupes de protéines effectuant les mêmes tâches peuvent être distingués et regroupés dans un cluster affecté à une fonction biologique reconnue pour ce module. Comme observé dans [90], une définition généralement acceptée de «cluster» n'existe pas dans le contexte des réseaux, car elle dépend du domaine d'application spécifique. Cependant, il est largement admis qu'une communauté devrait avoir plus de liens internes que les connexions externes [91].

2.2 Les Réseaux

2.2.1 Propriétés structurelles des réseaux

L'analyse des réseaux traite des propriétés statistiques qui caractérisent la structure de ceux-ci. Dans la suite je décris ces propriétés. Mais il convient initialement de décrire l'application des graphes à l'étude des réseaux.

Formellement un graphe $G = (V, E)$ est défini par un ensemble de nœuds V et une relation $E \rightarrow V \times V$ définissant quels nœuds sont connectés. Il existe deux façons classiques de représenter un graphe : par une matrice d'adjacence ou par un ensemble de listes d'adjacence.

Supposons que les sommets du graphe $G = (S, A)$ sont numérotés de 1 à n, avec $n = |S|$. La matrice d'adjacence d'un graphe est une matrice booléenne, appelé matrice d'adjacence, M de taille $n \times n$ telle que $M[i][j] = 1$ si $(i, j) \in A$, et $M[i][j] = 0$ sinon.

Un graphe valué associe une fonction v : $E \rightarrow L$ qui projette chaque lien du graphe dans un espace de label L. Dans ce cas, on indique la valeur d'un lien dans la matrice d'adjacence, par $M[i][j] = v(e_{ij})$.
Si la matrice M est symétrique, le graphe est non orienté, sinon il est orienté.

La seconde méthode de représentation d'un graphe est par liste d'adjacence. Cette méthode consiste en un tableau *T* de *n* listes, une pour chaque sommet dans *V*. Pour chaque sommet *s* la liste d'adjacence *T*[*s*] est une liste contenant tous les sommets adjacents à *S*.

2.2.2 Cheminements et connexités

2.2.2.1 Cheminements

Notions de chemin, chaîne, cycle et circuit

Dans un graphe orienté, un chemin d'un sommet *u* vers un sommet *v* est une séquence $<u = S_v, S_1, \dots, S_u = v>$, tels que $(S_v, S_{i+1}) \in E$ pour $i \in 0,\cdots,N$. Un sommet *v* est accessible de *u* s'il existe un chemin de *u* à *v*. Un chemin est élémentaire si les sommets qu'il contient sont tous distincts sinon c'est un chemin avec boucle. Un circuit est un chemin $< S_v, S_1, \dots, S_u>$ dont le début est l'extrémité sont le même nœud. Considérons par exemple (Figure 2.1) le graphe orienté suivant :

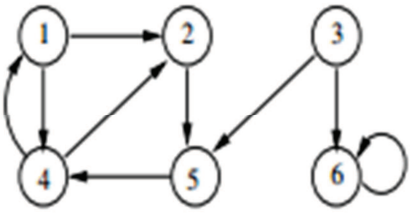


Figure 2.1 : Exemple d'un graphe orienté

- Un chemin élémentaire dans ce graphe est < 1, 4, 2, 5 >.
- Un chemin non élémentaire dans ce graphe est < 3, 6, 6, 6 >.
- Un circuit élémentaire dans ce graphe est < 1, 2, 5, 4, 1 >.
- Un circuit non élémentaire dans ce graphe est < 1, 2, 5, 4, 2, 5, 4, 1 >.

On retrouve ces différentes notions de cheminement dans les graphes non orientés. Dans ce cas, on parlera de chaîne au lieu de chemin, et de cycle au lieu de circuit. Un graphe sans cycle est dit acyclique.

2.2.2.2 Connexité

Un graphe non orienté est connexe si chaque sommet est accessible à partir de n'importe quel autre. Autrement dit, si pour tout couple de sommets distincts, il existe une chaîne entre eux. Par exemple, le graphe (Figure 2.2) non orienté suivant n'est pas connexe.

Figure 2.2 : Graphe non orienté

Il n'existe pas de chaîne entre les sommets *a* et *e*. En revanche, le sous-graphe défini par les sommets {*a, b, c, d*} est connexe.

Une composante connexe d'un graphe non orienté *G* est un sous-graphe G_0 qui est connexe et maximal, c'est à dire qu'aucun autre sous-graphe connexe de G ne contient G_0. Par exemple, le graphe précédent est composé de 2 composantes connexes : la première est le sous-graphe défini par les sommets {*a*, *b*, *c*, *d*} et la seconde est le sous-graphe défini par les sommets {*e*, *f*, *g*}.

Notons que si l'on calcule la fermeture transitive [92] d'un graphe connexe, on obtient un graphe complet, c'est-à-dire un graphe ou tout nœud et connecté à tous les autres. De même, si l'on calcule la fermeture transitive d'un graphe comportant *k* composantes connexes, on obtient un graphe contenant k sous-graphes complets (un pour chaque composante connexe). Par exemple (Figure 2.3), la fermeture transitive du graphe précédent est :

Figure 2.3 : Graphe connexe

On retrouve ces différentes notions de connexités dans les graphes orientés, en remplaçant naturellement la notion de chaîne par celle de chemin : on parle de graphe fortement connexe (Figure 2.4) au lieu de connexe, de composante fortement connexe au lieu de composante connexe.

Figure 2.4 : Graphe fortement connexe

Par contre le graphe orienté suivant contient 2 composantes fortement connexes (Figure 2.5) : la première est le sous-graphe défini par les sommets {*a, b, c, d*} et la seconde est le sous-graphe défini par les sommets {*e, f, g*}.

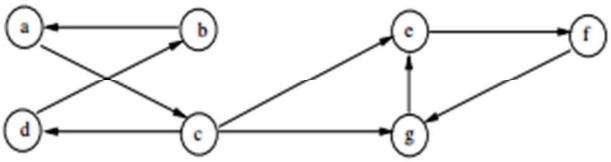


Figure 2.5 : Graphe orienté fortement connexe

Comme pour les graphes non orientés, une façon de déterminer si un graphe orienté est fortement connexe consiste à calculer sa fermeture transitive : si la fermeture transitive du graphe est le graphe complet, alors il est fortement connexe. Notons qu'il existe un algorithme bien plus efficace pour déterminer les composantes fortement connexes d'un graphe non orienté [92].

Dans la suite nous décrirons trois modèles de réseaux qui peuvent utile à la compréhension des réseaux d'acides aminés ou d'atomes. Mais avant nous devons décrire la distribution de degré dans les réseaux.

2.3 Distribution de degré des réseaux

Un graphe peut avoir une structure extrêmement complexe et les connexions entre nœuds peuvent présenter des motifs compliqués. Un défi dans l'étude des graphes complexes est de développer des mesures simples qui capturent des éléments importants de la structure de façon compréhensible. Une de ces mesures consiste à regarder chaque nœud séparément et à étudier, le degré du nœud, c'est à dire son nombre de connexions/liens avec ses voisins.

2.3.1 Graphes non orientés

Nous nous limitons pour le moment, à des graphes non orientés. Dans ces graphes le degré d'un nœud *i* est le nombre de voisins direct dont il dispose. En termes de la matrice d'adjacence A, le degré de nœud *i* est la somme de la ligne correspondante au nœud *i* dans A,

$$\mathbf{k_i = \sum_j a_{ij}}$$

La distribution de degré d'un graphe P_{deg} (k) est obtenue comptant la fraction de nœuds dans le graphe dont le degré est k. Je montre un petit exemple dans la figure 2.6 suivante :

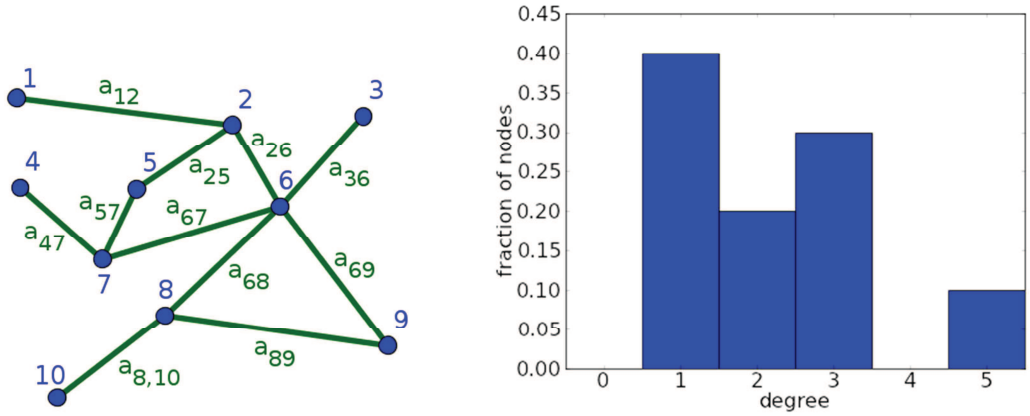


Figure 2.6 : a : Petit graphe non orienté avec des nœuds numérotés et étiquetés b : Distribution de Degré du graphe

La distribution des degrés ne donne clairement qu'une petite quantité d'informations sur un réseau. Mais cette information simple donne des indices importants sur la structure d'un réseau. Par exemple, dans un graphe complet où tous les nœuds sont connectés à tous les autres la distribution de degré est nulle sauf pour une seule valeur, n-1 ou n est le nombre de nœuds. Cependant, les graphes obtenus sur des réseaux du monde réel ont généralement des distributions de degrés différentes. Par exemple dans la plupart des graphes issus du monde réel, la plupart des nœuds ont un degré relativement faible, mais quelques nœuds ont un très grand degré. Par analogie avec les réseaux de transport, ces nœuds avec un grand degré sont souvent désignés comme des hubs, comme les aéroports de correspondance. Par exemple, dans le graphe dont la distribution de degré est présentée (Figure 2.7), le degré moyen est d'environ 7, mais 3/4 des nœuds ont un degré inférieur à 3. La moyenne est tirée vers le haut à 7 par la présence de quelques hubs avec des degrés de l'ordre d'une centaine.

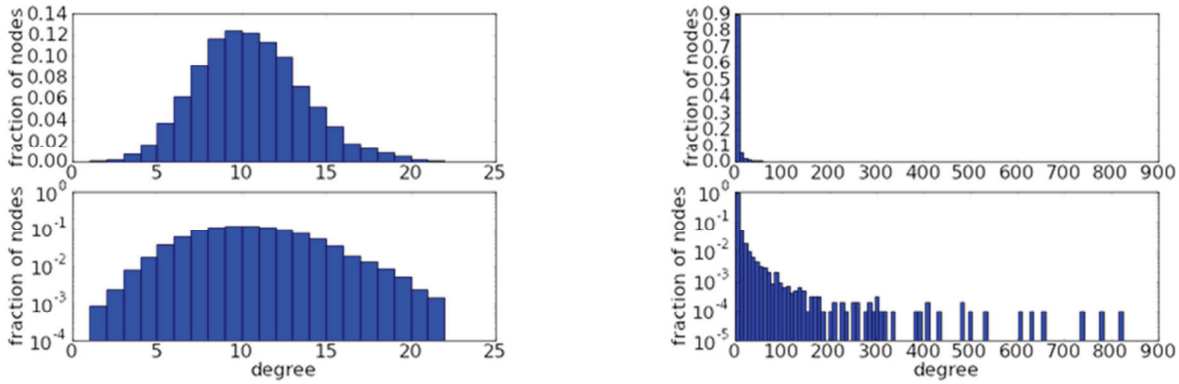


Figure 2.7 : a : Distribution de degré d'un réseau avec 10.000 nœuds, une distribution de degré binomiale avec un degré moyen de 10. b : Distribution de degré de d'un graphe avec 10.000 nœuds suivant une loi de puissance dont le degré moyen d'environ 7. Dans les deux figures l'histogramme du dessus est sur une échelle linéaire tandis que celui du bas montre les mêmes données sur une échelle logarithmique

Ces exemples illustrent le fait, qu'une mesure aussi simple que la distribution de degré peut nous donner une vision de la structure d'un réseau et de distinguer différents types de réseaux. Néanmoins, la distribution de degré ne capture pas la façon dont les nœuds sont reliés les uns aux autres.

2.3.2 Graphes orientés

La distribution de degré d'un graphe orienté est un peu plus complexe que pour les graphes non orientés. La raison est que le degré d'un nœud dans un réseau orienté ne peut être capturé par un seul nombre. En effet les liens peuvent être entrant ou sortant et en ignorant la direction des bords on perd une information importante. Ainsi il convient de compter séparément les liens entrants et sortants. Le degré entrant du nœud *i* est le nombre total de liaisons entrantes au nœud *i*, c'est-à-dire la somme de la ligne *i* de la matrice d'adjacence.

$$\mathbf{k^{in}_i = \sum_j a_{ij}.}$$

D'autre part, le degré sortant du nœud *i* est le nombre total de liaisons en provenance du nœud *i*, c'est-à-dire la somme de la colonne *i* de la matrice d'adjacence

$$\mathbf{k^{out}_i = \sum_j a_{ji}.}$$

Avec ces deux degrés, la distribution de degré devient une distribution bidimensionnelle, P_{deg} (k^{in}, k^{out}) = la fraction de nœuds dans le graphe avec degré entrant k^{in} et degré sortant k^{out} degré. On peut définir des distributions de degrés marginaux comme P^{in}_{deg} (k^{in}) = la fraction de nœuds dans le graphe avec degré entrant k^{in}. Ces distributions marginales

sont beaucoup plus simples à traiter. Mais l'étude séparée des distributions marginale suppose implicitement que l'on ne se préoccupe pas des corrélations entre les degrés entrant et sortant d'un nœud, alors que cette corrélation peut avoir un impact important sur les propriétés du graphe.

2.4 Groupes ou communautés : Notion du clustering

Le coefficient de clustering C est défini comme la probabilité moyenne que deux voisins d'un nœud soient adjacents. Pour un nœud *v* nous définissons le coefficient de clustering de ce nœud, $C_v \in [0,1]$, comme $C_v = 2E_v/d_v(d_v-1)$, où E_v est le nombre d'arêtes entre voisins de *v*, d_v est le nombre de voisins du nœud *v*. Le coefficient de clustering du réseau entier, *C,* est défini comme la moyenne des C_v, pour tous les nœuds du réseau.

Un ensemble de nœuds fortement interconnectés signifie généralement une forte similarité entre ces nœuds. Ainsi la recherche d'ensembles de nœuds qui sont très interconnectés entre eux, et une faible connectivité avec d'autres groupes de nœuds, ce qu'on appelle la recherche de communauté (Figure 2.8), est une des techniques les plus utilisées pour l'analyse exploratoire des données, avec des applications allant des statistiques, l'informatique, la biologie aux sciences sociales ou en psychologie. Je définirais plus loin plus précisément ce problème.

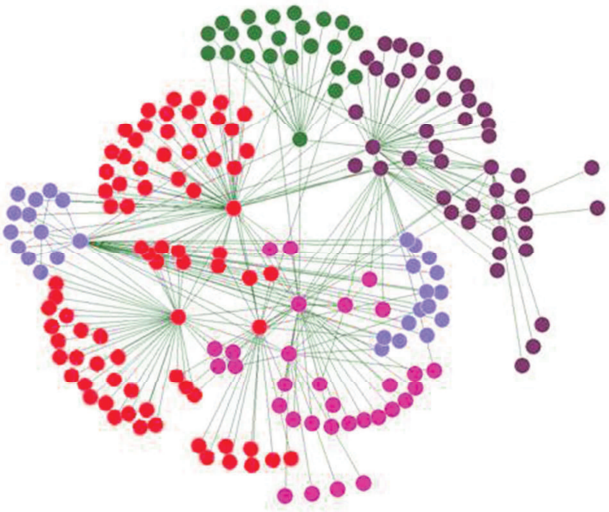


Figure 2.8 : Exemple d'une structure communautaire. Chaque couleur présente une communauté de protéines.

2.5 Robustesse fonctionnelle et dynamique

Les observations sur lesquelles les graphes sont construits, peuvent être sensibles à des événements qui aboutissent à la disparition ou l'ajout de nœuds, ou à la disparition où l'ajout de liens, *e.g.*, des mutations dans les acides aminés qui aboutissent à des changements sur les graphes de connectivités des atomes. L'effet d'une perturbation ne peut pas seulement dépendre du degré des nœuds. Bien évidemment l'effet du changement sur un nœud est liée à la fonction de ce nœud, mais des variations locales peuvent aussi avoir un impact important sur les propriétés globales du graphe. Par exemple, un graphe qui reste connexe grâce à un unique lien se retrouvera séparé en deux sous-graphes non connectés si ce lien particulier est supprimé. Il est donc pertinent d'étudier la robustesse d'un graphe à des perturbations qui pourrait arriver et changer la configuration locale de certains nœuds.

Il est néanmoins impossible d'étudier dans l'absolu la robustesse de tous les paramètres du graphe, et il faut se concentrer sur les paramètres qui ont une importance particulière pour l'objet sous-jacent d'où est issue le graphe. Par exemple, la plus courte distance entre deux nœuds ou le chemin entre deux nœuds est fréquemment un paramètre important. Le paramètre relatif à ceci dans le graphe est la "Betweenness centrality" des nœuds.

La "betweenness centrality" (*BC*) d'un nœud *i* est le nombre de plus courts chemins, appelé chemin géodésiques, entre deux nœuds du graphe qui passent par *i*. Dans la figure 2.9, j'illustre deux graphes dans lequel la *BC* du nœud *i* est de l'ordre de N^2 où *N* est le nombre de nœuds du graphe. Plus la valeur de la *BC* d'un nœud est grande, plus la suppression de ce nœud impacte de chemins dans le graphe et ainsi ce nœud a une grande importance dans la robustesse du graphe. Le diamètre d'un graphe étant défini comme le plus long des plus courts chemins entre tous les nœuds, un graphe peut être considéré comme robuste si la suppression des nœuds dont la *BC* est grande ne change pas le diamètre du graphe.

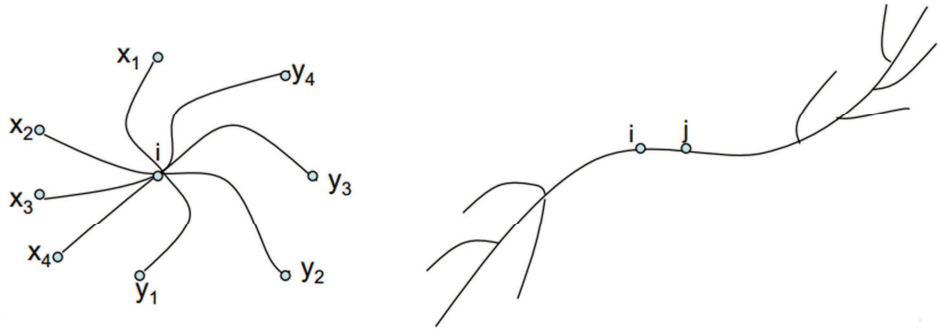


Figure 2.9 : Deux graphes où la *BC* du nœud *i est* grande.

2.6 Modèles de réseaux

2.6.1 Graphes aléatoires

Les graphes aléatoires (Figure 2.10) sont des graphes qui sont générés par un processus aléatoire. Ces graphes ont été introduits par Paul Erdös et Alfréd Rényi en 1959 [93] afin de prouver certains résultats sur les graphes. Plusieurs modèles de graphes aléatoires sont défini avec des propriétés différentes selon le modèle choisi [94].

Le modèle d'Erdös-Rényi consiste à considérer un graphe à n sommets où l'existence de chaque arc avec une probabilité p est indépendante de celle des autres [94]. Ces graphes sont généralement notés G (n; p) où n est le nombre de sommets et p la probabilité que les arêtes soient présentes. En utilisant cette construction on obtient un graphe avec une distribution des degrés suivant une loi binomiale qui converge asymptotiquement vers une distribution de Poisson. Néanmoins, la majorité des graphes dans la nature ne suivent pas ce modèle [95].

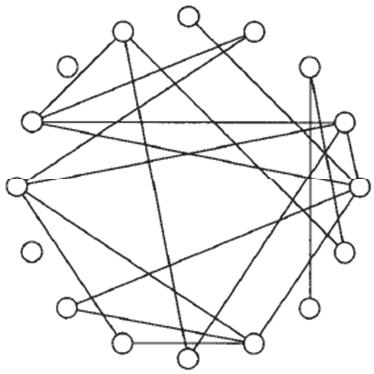


Figure 2.10 : Graphe aléatoire avec 16 sommets et 19 liens [96].

C'est pourquoi le modèle d'Erdös-Rényi a été étendu afin de construire des distributions des degrés définies. En particulier, des distributions de degrés suivant des lois de puissance sont souhaité afin d'obtenir des graphes sans échelle (*scale-free*) qui seront défini dans la suite.

2.6.2 Graphe petit monde (small-world)

Le modèle de graphe « petit monde » (*small world*) est issu des travaux de Watts et Strogatz. Ces graphes sont motivés par les travaux de Stanley Milgram concernant les fameux « six degrés de séparation » [97]. Ces graphes sont caractérisés par une distance moyenne entre deux nœuds qui est proportionnelle au logarithme du nombre de nœuds, et d'un grand nombre de sous-graphes qui sont proches de cliques, c'est-à-dire à peu près tous les nœuds de ces sous-graphes sont connectés entre eux.

Un graphe petit monde possède un coefficient de clustering très supérieur à un graphe aléatoire de même ordre (avec le même nombre de nœuds) et de même taille (avec le même nombre de liens) et un diamètre inférieur.

Le modèle petit monde est motivé par l'observation que de nombreux réseaux du monde réel présentent les deux propriétés suivantes :

- L'effet petit monde, c'est-à-dire la plupart des paires de sommets sont reliés par un court trajet à travers le réseau. Au cours des dernières années, le terme «effet petit monde » est venu à signifier expressément que la moyenne (ou parfois le maximum) de la distance entre nœuds dans le graphe augmente de façon logarithmique (ou plus lentement) avec le nombre total de sommets dans les réseaux. Ce genre de croissance logarithmique existe empiriquement dans les graphes issus de réseaux du monde réel. Watts et Strogatz ont implicitement assumé cette échelle en définissant un trajet court dans un graphe comme un trajet dont la longueur est comparable à celles observées dans un graphe aléatoire de même dimension et de même degré moyen. La longueur moyenne de trajet dans un graphe aléatoire et sa relation avec la taille de graphe sont connus analytiquement.
- Fort "clustering" ou "transitivité", ce qui signifie qu'il y a une forte probabilité que deux sommets soient connectés directement s'ils ont un autre voisin en commun [96].

La plupart des graphes issus de données empiriques, correspondent à la définition de graphes petit monde [98]. Ainsi, ces graphes, bien que d'essence théorique, se retrouvent empiriquement dans la nature contrairement aux graphes réguliers.

La figure 2.11 ci-dessous résume la méthode de construction de ces graphes. On commence par un graphe *k*-régulier, où chaque nœud est connecté à *k* autres nœuds, ou tous les sommets ont le même degré *k*. On supprime ensuite de façon aléatoire, un lien et on ajoute un lien au graphe de départ. Progressivement, le graphe *k*-régulier est transformé en graphe. Entre ces deux extrêmes, le graphe produit présente des propriétés petit-monde.

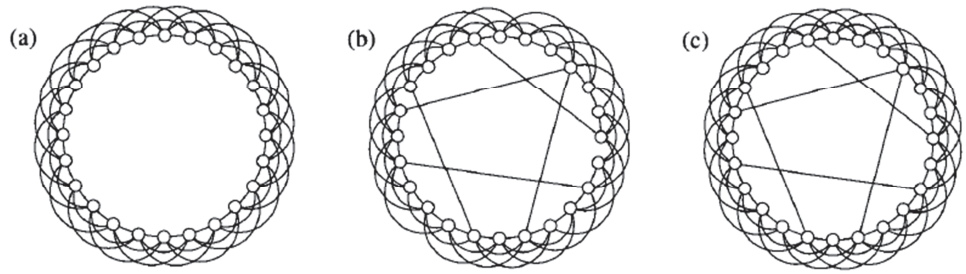


Figure 2.11 : (a) : le graphe initial *k*-régulier (b) : le graphe après qu'une petite fraction *p* des nœuds de (a) sont "refait" en déplaçant l'une de leurs liens vers un nouveau sommet choisi aléatoirement. (c) : variante de la construction dans laquelle des bornes de "coupure" sont ajoutés entre des paires de nœuds sélectionnés de manière aléatoire dans (a), et qu'un lien aléatoire est ajoutée [96].

2.6.3 Les graphes sans échelle (*scale-free)*

Barabàsi et Albert ont introduit en 1999 les graphes sans échelles ou *scale-free*[87] (Figure 2.12). Un graphe est dit invariant d'échelle (scale-free) si sa distribution de degré suit une équation $\boldsymbol{P(k) \sim k^{-\gamma}}$, avec $\gamma > 2$ [98]. Ainsi ces graphes présentent une distribution de degré en loi de puissance. Le terme sans échelle signifie ici que les nœuds n'ont pas un degré caractéristique car ils couvrent une trop large gamme de degré.

Ce modèle de graphes est intéressant, car contrairement aux deux autres modèles présentés précédemment, c'est-à-dire, aléatoire et petit monde, il ne repose pas sur une construction a priori, mais sur une analyse empirique de graphes existants. Ces graphes sont générés par des méthodes d'attachement préférentiel, ou un nœud va se connecter à un autre nœud en fonction de son degré donnant une vision assez hiérarchique du graphe.

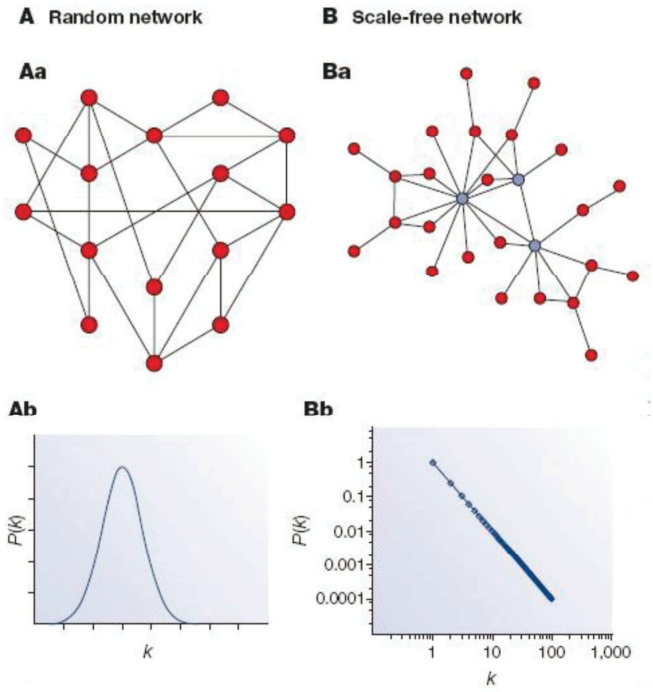


Figure 2.12 : Un réseau scale-free (Ba) suit une distribution en loi de puissance des degrés (Bb) et non une loi binomiale (Ab) comme les graphes aléatoires (Aa) [99].

2.7 Clustering spectral

Comme nous l'avons décrit précédemment, la recherche de communautés dans un graphe est une des techniques les plus utilisées pour l'analyse exploratoire des données, avec des applications allant des statistiques, l'informatique, la biologie aux sciences sociales ou en psychologie. Nous décrivions dans cette section précisément une des principales techniques permettant ceci.

2.7.1 Laplacien d'un graphe

Soit $G = (V, E)$ un graphe avec $| V | = n$ nœuds. Nous définissons la matrice n x n d'adjacence avec poids W du graphe, comme $W = (w_{ij})$, où w_{ij} est le poids du lien entre le nœud i et j et $w_{ij}=0$ signifie que les nœuds i et j ne sont pas connecté par un lien. Nous définissons le poids d'un nœud i comme suit :

$$d_i = \sum_j w_{ij}$$

Et D, la matrice diagonale ayant les poids des nœuds sur sa diagonale. Nous définissons aussi le $W(A,B)$ comme suit :

$$W(A,B) = \sum_{i \in A, j \in B} w_{ij}$$

Le poids reliant l'ensemble de nœuds *A* l'ensemble *B*.

Les définitions précédentes permettent de définir le Laplacien *L* d'un graphe comme suit :

$$L = D - W$$

Par construction le Laplacien du graphe est une matrice symétrique positive semi définie. Elle admet donc une décomposition spectrale. Etant donné que la somme des lignes de *L* est nulle, L admet une valeur propre égale à 0, avec un vecteur propre correspondent qui est le vecteur constant 1. Le Laplacien possède *n*-1 autres valeurs propres non-négatives $0 = \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$.

Supposez que les nœuds contiennent des valeurs f_i et que le vecteur *f* contient toutes ces valeurs. L'intérêt du Laplacien est défini par la propriété suivante :

$$f^T L f = \sum_{i,j \in V} w_{ij} (f_i - f_j)^2$$

Ainsi si c'est le graphe qui par le biais des interactions aboutit à une distribution des valeurs sur les nœuds définis par le vecteur *f*, le Laplacien définit la matrice d'interactions.

Le concept de Laplacien d'un graphe peut être étendu au Laplacien normalisé, qui normalise les éléments du Laplacien par rapport aux variations de degré être les nœuds. Le Laplacien normalisé est défini comme suit :

$$L^* = D^{-1/2} L D^{-1/2} = I - D^{-1/2} W D^{-1/2}$$

2.7.1 Coupe dans un graphe

Le problème de détection de communauté se réduit à partitionner l'ensemble des nœuds en *k* sous-ensembles, chacun contenant une des communautés que nous recherchons. Mais toutes les partitions ne sont pas équivalentes. Nous utilisons une fonction de cout la mesure $cut(A_1, A_2, \ldots, A_k)$ pour évaluer une partition $A_1, A_2, \ldots, A_k$.

$$cut(A_1, A_2, \ldots, A_k) = \frac{1}{2} \sum_{i=1}^{k} W(A_i, \overline{A_i})$$

où *W*(*A*,*B*) et $\overline{A_i}$ est le complémentaire de l'ensemble A_i. Cette fonction mesure le poids des liens qu'il faut couper afin de séparer l'ensemble des nœuds en communautés $A_1, A_2, \ldots, A_k$.

Le problème classique du *mincut* dans les graphes consiste à trouver une partition du graphe en k sous-graphes séparés qui minimiserait la valeur de la mesure $cut(A_1, A_2, \dots, A_k)$. Néanmoins, la résolution de ce problème, ne résout pas notre problème de détection de communauté. En effet, la résolution de ce problème aboutit généralement à des communautés de très petite taille (avec seulement un seul nœud). Ainsi on souhaite que les communautés $A_1, A_2, \dots, A_k$ soit suffisamment grandes. Ainsi nous remplaçons la fonction de cout $cut(A_1, A_2, \dots, A_k)$ par la fonction de cout de coupe normalisée suivante :

$$Ncut(A_1, A_2, \dots, A_k) = \frac{1}{2}\sum_{i=1}^{k} \frac{W(A_i, \overline{A_i})}{vol(A_i)}$$

où $vol(A_i)$ indique le volume de la partie A_i. Avec cette nouvelle fonction de cout une partition trop petite aura une valeur grande. Ainsi si nous trouvons la partition qui minimise la valeur de $Ncut(A_1, A_2, \dots, A_k)$ nous trouverons des parties bien connectés à l'intérieur, puisque leur volume sera grand, et faiblement connectés à l'extérieur puisque la valeur de $W(A_i, \overline{A_i})$ sera faible.

Ainsi le problème de détection de communautés qui a la base semblait difficile à définir se réduit à la résolution d'un problème d'optimisation sur un graphe. La technique de clustering spectrale permet de trouver une solution approximative au problème précédent et ainsi de trouver des solutions au problème de détection de communauté.

2.7.2 Algorithme de clustering spectral

Etudions la signification intuitive d'une décomposition spectrale, autrement dit une diagonalisation, de la matrice du Laplacien normalisé. Cette décomposition donne en sus des *n* valeurs propres de la matrice, un ensemble de *n* vecteurs propres qui donnent une base canonique et orthogonale pour l'espace d'interaction du graphe. On peut donc approximer de façon optimale (au sens du minimum d'erreur carrée) l'espace des interactions des nœuds du graphe en utilisant un sous-ensemble des *k* vecteurs propres correspondant aux valeurs propres les plus grandes.

Cette projection de l'espace d'interaction sur le graphe dans l'espace canonique permet aussi de représenter le graphe par un ensemble de *n* points dans un espace métrique, où chaque point représente un nœud du graphe, c'est à dire, le point *i* est défini en prenant les $i^{ème}$ coordonnées des *k* vecteurs propres choisis. Ainsi le graphe qui est un objet discret, se représente dans un espace métrique. La proximité des points dans ce nouvel espace métrique, signifie que les nœuds correspondants à ces points ont une interaction forte par le biais du

graphe. Ainsi en regroupant les points proches dans ce nouvel espace métrique nous trouvons des sous-ensembles de nœuds qui interagissent fortement, et quand nous mettons ces sous-ensembles le plus loin possible l'un de l'autre, nous nous assurons qu'ils ont une interaction faible entre eux.

Le clustering spectral consiste en deux étapes distinctes. Dans une première étape nous diagonalisons la matrice du Laplacien normalisée afin de trouver les k vecteurs propres relatifs aux valeurs propres les plus grandes. Ceci nous permet de trouver les n points représentants les n nœuds du graphe. Dans une seconde étape nous appliquons un algorithme de k-means afin de regrouper les points dans l'espace métriques des points. Les partitions obtenues par le k-means sont les communautés que nous recherchons.

2.7.3 Application à l'analyse des réseaux inter-atomes dans les protéines

Dans la suite, je décrirai comment les éléments précédents ont été appliqués dans notre cas d'étude grâce à un logiciel développé dans notre équipe qui prend le nom Spectral-Pro [83].

L'information initiale de Spectral-Pro est le fichier au format PDB décrivant une protéine. Ce fichier contient les coordonnées 3D des différents atomes de la protéine ainsi que l'appartenance de ces atomes à différentes chaînes constituant la protéine. La première étape du consiste à traiter le fichier PDB afin de construire un graphe de connectivité entre les atomes. Nous supposons deux atomes connectés s'ils appartiennent au backbone de deux chaînes différentes et si leur distance est inférieure à 5Å.

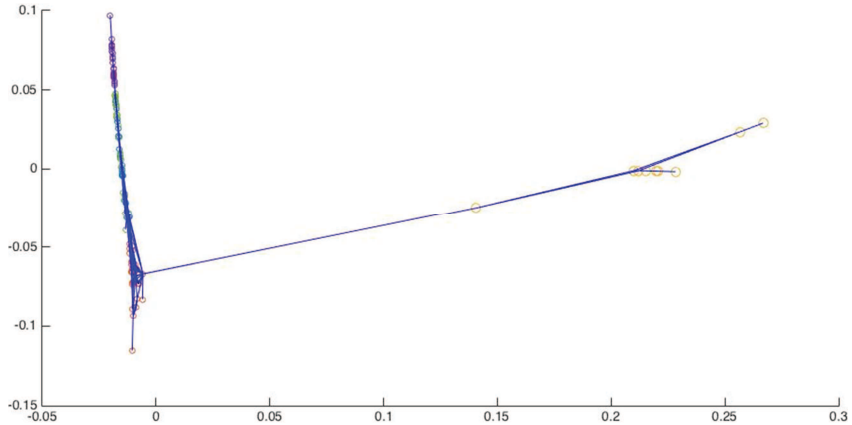


Figure 2.13 : Représentation du graphe dans l'espace métrique résultant de la diagonalisation du Laplacien pour la protéine 1EEI

Ce graphe de connectivité n'est généralement pas connexe. Nous devons donc le séparer en composante connexe. Le reste de l'action de Spectral-pro s'applique à chaque composante connexe séparément. Spectral-pro applique à chaque composante connexe un clustering spectral comme décrit précédemment. Ce clustering sépare chaque composante connexe en un nombre de communautés donné *a priori*. Dans la suite, ces communautés sont affichées graphiquement. Le graphe entre les atomes, peut être étendu à un graphe entre les acides aminés où deux acides aminés sont connectés si au moins un atome de chacun de ces acides est connecté dans le graphe défini précédemment

Je présente dans les figures 2.13 et 2.14 l'application du Spectral-pro et du clustering spectral à la protéine 1EEI. Rappelons que la première étape dans le clustering spectral consiste à projeter les nœuds du graphe dans un espace métrique ou les nœuds interagissant fortement se retrouvent proches l'un des autres. Il suffit ensuite de regrouper les points proche dans cet espace par le biais d'un algorithme de clustering simple comme le *k*-means afin d'obtenir le clustering spectral. La figure 2.13 montre une représentation en deux dimensions de cet espace métrique ainsi que le résultat du clustering spectral. Les nœuds de même couleur appartiennent au même cluster. La figure 2.14 représente le graphe entre les acides aminés où les acides aminés sont coloriés de la même couleur que le cluster auquel il appartient.

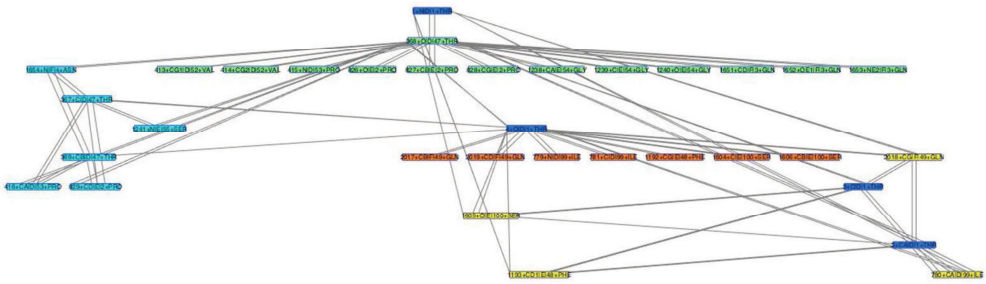


Figure 2.14 : Communautés d'acides aminés détectés par le clustering spectral

La figure 2.14 montre l'intérêt principal du clustering spectral et de Spectral-pro dans l'analyse de la structure des protéines. En effet cette représentation permet de détecter rapidement les interfaces qui sont positionnées au niveau des liens entre les nœuds de couleurs différentes et de visualiser simplement les relations entre les acides aminés et de détecter ceux qui pourraient potentiellement avoir un intérêt biologique. Le chapitre suivant décrira en détails la méthodologie d'analyse de la structure des protéines en utilisant Spectral-pro.

Chapitre 3: Méthodologie

Mes travaux de recherche ont porté sur l'étude des propriétés des réseaux d'interactions entre atomes des acides aminés localisés aux interfaces entre plusieurs chaînes protéiques. Mon travail a donc été particulièrement focalisé sur l'étude des protéines oligomériques qui sont composées de plusieurs chaînes protéiques associées entre elles et sur la modélisation des interfaces protéiques par des réseaux. Pour mes travaux, j'ai utilisé des outils informatiques que je décrirai dans la suite.

L'outil informatique a été utilisé dans ma recherche, et plus largement en biologie, dans trois contextes qui sont fréquemment complémentaires et qui coexistent.

Un premier contexte d'utilisation de l'outil informatique est le traitement, la représentation et la visualisation de l'information biologique. C'est dans ce contexte que se situe des logiciels comme *jmol* ou *pymol* qui permettent une visualisation de la structure tridimensionnelle des protéines, des logiciels comme *dotviewer* ou *Gephi* qui permettent la visualisation de graphes (réseaux) ou même les outils de bureautique.

Un second contexte est la simulation. Le travail expérimental *in vitro* et *in vivo* sur les protéines est difficile, et pour certaines questions il peut même être inadapté dans l'état actuel des technologies. Par exemple, l'ensemble des conformations (structures) qu'une protéine prend pour répondre à sa fonction ne peut pas être établi expérimentalement car certaines conformations ne seront jamais viables ou détectables expérimentalement. Il faut appréhender le problème par des approches théoriques. Ces dernières années les avancées en biologie moléculaire et en biophysique ont permis de produire une grande quantité de données (ex. le nombre de structure atomiques) exploitables par l'informatique pour proposer des modèles mathématiques de protéines suffisamment fiables pour reproduire le comportement des protéines. Ces outils de modélisation permettent parfois d'exécuter des expériences virtuelles sans avoir besoin de déployer un protocole expérimental complexe. Le logiciel *Fold-X* que je décrirai dans la suite est un exemple de ce type d'outil logiciel. D'autres outils fondés sur des modèles chimie physique rentrent aussi dans ce contexte. En parallèle, les progrès en puissance de calculs ont rendu possible le traitement de ces grands nombres de données, permettant aux approches de biologie informatique leur essor actuel.

Le dernier contexte, qui est fondamental dans cette thèse, est la fouille de données. Il convient ici de clarifier notre utilisation de la fouille de données. L'objectif des méthodes de fouilles de données est de rechercher dans des données mises en forme (i,e. mises sous forme de graphe, ou de tableau, etc.), des corrélations ou des cooccurrences non triviales. Ces corrélations et/ou cooccurrences doivent ensuite être considérées comme candidates à des investigations plus poussées permettant de valider leur pertinence biologique. Du fait de l'importance de cette validation, je présente dans la suite une section méthodologique.

3.1 Méthodologie de l'application de la fouille de données aux problèmes biologiques

Dans cette thèse comme dans la plupart des travaux de recherche en biologie, il y a des observations que nous souhaitons interpréter, par exemple ici nous observons la structure tridimensionnelle des protéines dans le but d'en comprendre la dynamique. Mais souvent ces observations sont complexes et nos connaissances *a priori*, des mécanismes ayant abouti à ce qu'on observe sont aussi parcellaires. L'idée des approches de fouille de données est de rechercher dans les données issues de l'observation des similarités ou des régularités avec deux objectifs liés : d'une part simplifier la description de l'observation en la catégorisant en un petit nombre de composantes contenant des observations liés, d'autre part interpréter l'observation en s'appuyant sur ces composantes afin de trouver une explication, ou une cause, à ce que l'on observe. Ainsi la fouille de données consiste en trois étapes : définition d'une mesure de la liaison entre similarité ou régularité, regroupement des observations similaires et interprétation.

Ce qui fait la force de la fouille de données et qui explique sa large utilisation est liée au fait que les deux premières étapes décrites plus haut peuvent être réduites à des briques logicielles génériques qui peuvent être implantées dans des logiciels de fouilles de données. Ainsi il existe plusieurs mesures de similarité qui sont largement utilisées. Par exemple, les distances euclidiennes ou euclidiennes pondérées sont utilisées, quand les données sont multidimensionnelles avec un nombre de paramètres limités et fixes. Ces mesures sont particulièrement attractives car elles permettent fréquemment l'application de l'arsenal des outils de la statistique normale par le biais d'une interprétation des similarités moyennes sous forme de variances de distribution gaussienne. Un autre exemple important est le cas où la mesure de similarité est binaire, deux observations sont similaires ou non. Dans ce cas cette

similitude est décrite par une relation qui définit un graphe sur les observations. Une mesure de similitude peut être transformée en graphe en utilisant un seuil de connectivité.

La définition d'une mesure de similarité ou d'un graphe entre les éléments d'observations ouvre la voie à l'utilisation de méthodes génériques permettant de regrouper les observations semblables. Il convient néanmoins avant d'aller plus loin de se poser la question de la pertinence d'une mesure de similarité ou d'une relation de graphes. Bien évidemment une mesure de similarité ou une association peut résulter d'éléments théoriques ou d'informations externes que nous avons, mais en pratique il est difficile de connaître *a priori* la pertinence d'une mesure de similarité ou d'une relation en particulier pour des observations complexes. Ainsi, la mesure de similarité ou la relation utilisée pour relier les observations entre elles, est hypothétique en ce sens que nous ne sommes pas assurés que ce soit une mesure pertinente pour l'application finale. Ce problème est encore plus important quand c'est un graphe qui capture la similarité car la décision de lier deux observations est fondée sur un seuil difficile à définir précisément. Ainsi, la fouille de données a un troisième objectif qui est plus rarement pris en compte : décider de la pertinence d'une liaison prévue par la mesure de similarité. Il est à noter que la pertinence de la mesure utilisée ne peut être décidée qu'après coup, c'est à dire après avoir vérifié si le regroupement obtenu à l'issue de la fouille de données est pertinent par rapport à l'application finale. Ainsi la fouille de données apparaît fréquemment comme un processus itératif où une mesure candidate est utilisée pour faire un regroupement utile pour affiner la mesure de similarité correspondante.

Après l'étape de mesure de similarité, l'étape de regroupement des observations similaires est principalement algorithmique et dépend de considérations liées à la complexité de l'opération de regroupement. Bien évidemment l'algorithme précis à utiliser dépend de la mesure de similarité choisie à la première étape. Il convient d'apporter des adaptions afin de prendre en compte la complexité numérique. Par exemple le nombre de dimensions de la mesure de similarité a généralement un impact direct sur la complexité de l'opération de regroupement. Il convient ainsi de réduire le nombre de dimensions utilisé dans la mesure de similarité afin de simplifier le traitement. Il conviendra aussi de choisir la bonne heuristique permettant en même temps de réduire la complexité du calcul tout en s'assurant que les hypothèses de l'heuristique restent compatibles avec les conditions de nos observations.

La dernière étape de la fouille de données qui est aussi la plus importante est l'interprétation. Cette étape permet de revenir du monde abstrait et informatique de la fouille de données au monde concret de l'application biologique qui nous intéresse. Il est notable que le regroupement obtenu par fouille de données peut être pertinent par rapport au problème étudié ou n'être qu'un artéfact de la mesure de similarité utilisée ou de la configuration particulière des observations. C'est notre rôle de filtrer les résultats de la fouille de données afin de détecter ceux qui ont un sens biologique. Ainsi il est très difficile d'étendre aveuglement des interprétations faites sur un ensemble de données sur un autre ensemble différent. Malheureusement, la littérature actuelle est remplie d'exemples où des inférences effectuées par exemple sur les réseaux sociaux sont directement étendues aux réseaux biologiques sans analyse critique.

La discussion précédente nous permet d'ébaucher les bases méthodologiques de notre utilisation de la fouille de données dans cette thèse. La donnée brute que nous utilisons dans cette thèse est la structure 3D de protéines. Cette donnée brute est transformée en mesure de proximité, que nous décrirons plus tard dans ce chapitre. Ces mesures de proximité sont transformées en graphes qui définissent les interactions possibles/potentielles entre atomes des différentes chaînes qui constituent une protéine oligomérique. Nous appliquons des techniques de fouilles de données sur les graphes afin de détecter des regroupements d'atomes. Ces regroupements permettent de filtrer entre les liens matérialisant les relations entre atomes, ceux qui sont pertinents pour comprendre la structure protéique de ceux qui ne sont peut-être pas importants. Ces regroupements sont interprétés par le biais des interfaces protéiques. Nous validons la pertinence et la valeur de l'interprétation par une expérience consistant à faire muter les acides aminées participant aux interfaces que nous avons détecté et à évaluer si ces mutations aboutissent à une restructuration importante de la protéine et éventuellement à des cas de mutations protéiques pathogènes. Ainsi, la fouille de données est utilisée dans mes travaux comme un filtre permettant de générer, entre toutes les hypothèses possibles, un nombre réduit d'hypothèses plausibles de mutations aboutissant à des effets importants sur la structure protéique. Ce faisant, la fouille de données me permettra d'avancer dans la compréhension du processus d'assemblage protéiques et plus généralement l'émergence des structures protéiques.

3.2 Description des outils utilisés

Dans la suite, je présente les outils essentiels que j'ai utilisés durant ma thèse. Je commence par décrire les deux outils de fouille de données que j'ai utilisées. Le premier qui s'appelle Gemini, a été développé par Giovanni FEVERATI. Il permet d'examiner les propriétés des interfaces des protéines oligomériques. Le second développé par Kavé SALAMATIAN s'appelle Spectral-Pro. Il utilise des techniques d'analyse spectrale des graphes pour effectuer un clustering de graphe. Ces deux outils ont permis selon la description de la section précédente, d'obtenir des hypothèses biologiques à investiguer plus profondément.

Je décris ensuite Fold-X, un outil de simulation permettant d'obtenir une estimation rapide et quantitative de l'importance des interactions qui contribuent à la stabilité des protéines et des complexes protéiques.

3.2.1 Programme Gemini

3.2.1.1 Définition

Le logiciel Gemini a été développé par Giovanni FEVERATI au laboratoire de Physique Théorique de l'université de Savoie (LAPTH) dans le cadre d'un projet interdisciplinaire soutenu par la Fédération de recherche, FR2914 MSIF (Modélisation, Simulations, Interactions Fondamentales). Ce logiciel isole à partir des coordonnées cartésiennes des atomes d'une protéine, fournies par la PDB, les acides aminés impliqués dans l'interface d'un oligomère. Gemini plutôt qu'identifier tous les hots spots d'une interface, extrait le sous-ensemble le plus petit caractérisant "les interactions intermoléculaires".

Gemini permet de proposer un squelette d'interactions entre les acides aminés impliqués dans une interface prenant en compte leur rôle dans la spécificité de l'interface et dans la régulation du mécanisme d'assemblage. En particulier, Gemini permet de comparer les interfaces protéiques de géométrie similaire [100].

Gemini est constitué d'une série de programmes permettant d'étudier différentes propriétés des interfaces de l'oligomère: GeminiDistances, GeminiRegions, GeminiGraph et GeminiData. Le schéma de principe de Gemini est indiqué en figure 3.1.

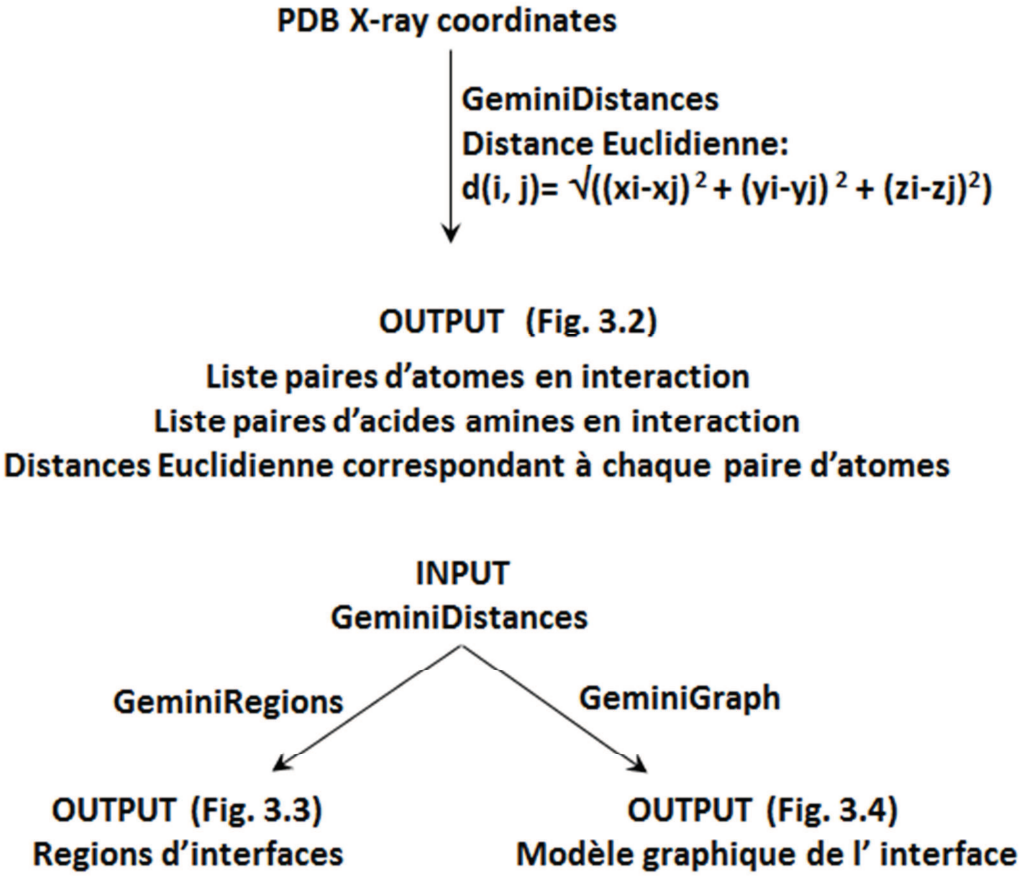


Figure 3.1 : Schéma de principe de Gemini

3.2.1.2 Gemini Distances

Alors qu'il est difficile de définir manuellement les interfaces par simple visualisation, ce programme a pour objectif principal de détecter l'interface entre deux chaînes voisines *M* et *M* +1 dans une protéine oligomérique.

Toutes les liaisons chimiques faibles se produisent à une distance inférieure ou égale à 5Å. Donc le premier critère de sélection est de ne conserver que les couples d'atomes se situant à une distance inférieure ou égale à 5Å l'un de l'autre. Comme il existait un trop grand nombre de couples d'atomes répondant à ce critère, une symétrisation est effectuée, pour désélectionner le plus grand nombre possible d'atomes en interaction. L'idée étant de ne conserver qu'un squelette minimum d'interaction.

La symétrisation choisit parmi toutes les paires d'atomes celles dont les atomes sont mutuellement les plus proches. En d'autres termes, si la distance AB entre l'atome A de la chaîne M et l'atome B de la chaîne M+1 est la plus petite pour A et B, Gemini conserve la

paire d'atomes (A, B), si B a un atome plus proche que A sur la chaîne M, alors la distance AB est éliminée. La symétrisation se fonde sur la plus grande probabilité d'avoir une interaction chimique entre atomes les plus proches. En effet des atomes plus éloignés seront moins surs de former une interaction du fait de l'effet d'écran produit par leurs voisins plus proches. Cette procédure permet en plus de ne pas sélectionner les paires d'atomes en interaction à partir d'une distance seuil et donc le réseau d'acides aminés construit reste le même sur de large conditions cristallines, comme par exemple autour des 5 Å et donc de l'erreur expérimental associée aux rayons X ~1 Å. Gemini Distances n'effectue aucune sélection chimique sur les paires parce qu'il est supposé que la solution cristalline est une solution « chimique » correcte, qui ne positionne pas à proximité des atomes chimiquement ou géométriquement (stériquement) incompatibles.

Gemini Distances génère un fichier de sortie regroupant les couples d'atomes en interaction, les acides aminés correspondant, et diverses autres informations sur la protéine (figure 3.2).

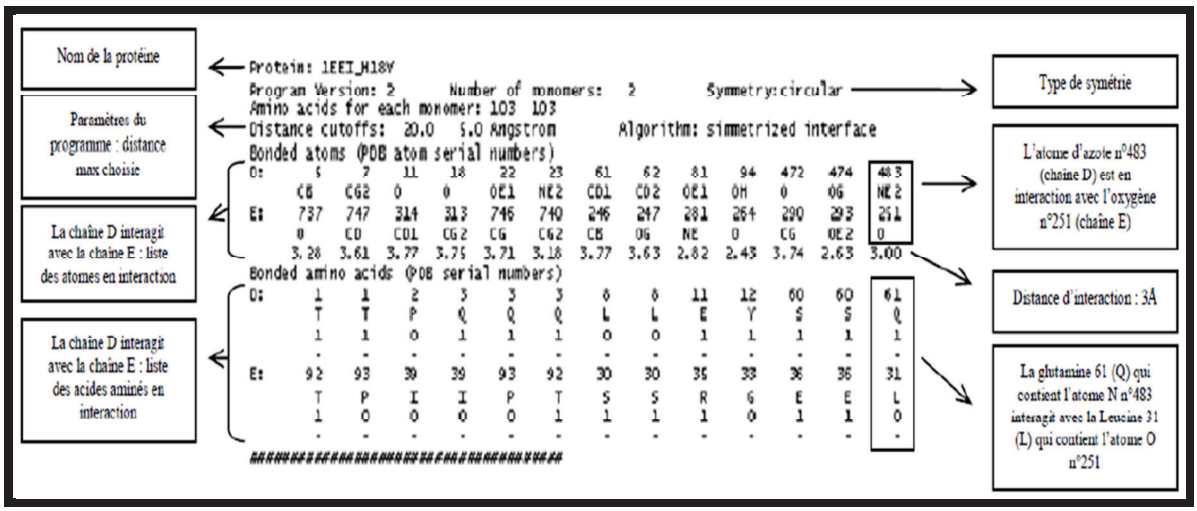


Figure 3.2: Fichier de sortie de GeminiDistances

3.2.1.3 Gemini Région

Ce programme sépare l'interface entre les acides aminés des deux chaînes adjacentes, extraite par Gemini Distances, en régions (Figure 3.3), appelés réseaux d'interactions élémentaires. Un segment constitué d'acides aminés contigus interagissant avec deux groupes d'acides aminés éloignés de plus de 5 acides aminés sur le segment opposé sera découpés en deux segments composant deux régions d'interfaces distinctes.

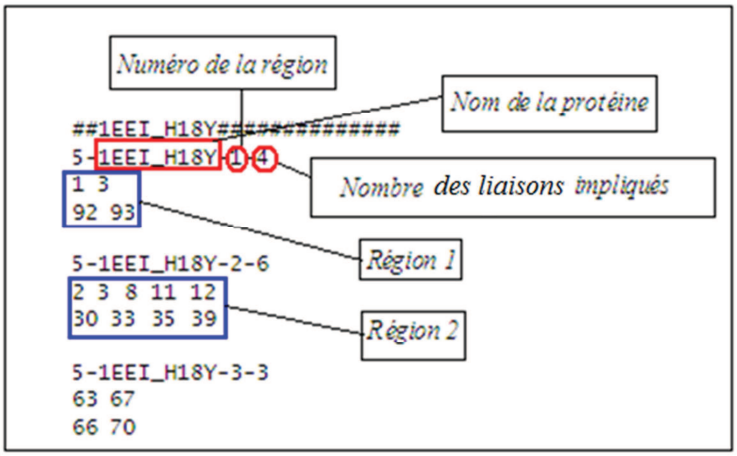


Figure 3.3: Exemple des régions d'interface données par Gemini Régions

3.2.1.4 Gemini Graphe

Ce programme crée une représentation graphique de l'interface des régions et de leurs interactions acides aminés sous forme d'un graphe où les sommets sont des acides aminés et les liens sont ceux détectés par Gemini Distance. Contrairement à Gemini Région et Gemini Distances, Gemini Graphe représente tous les acides aminés des segments et pas seulement ceux choisis comme ayant une interaction chimique (Figure 3.4). Ainsi les liens impliqués dans une liaison chimique faible entre acides aminés de l'interface sont représentés par une croix «X», alors que ceux qui ne sont pas impliqués dans de telles liaisons sont représentés par un point ".".

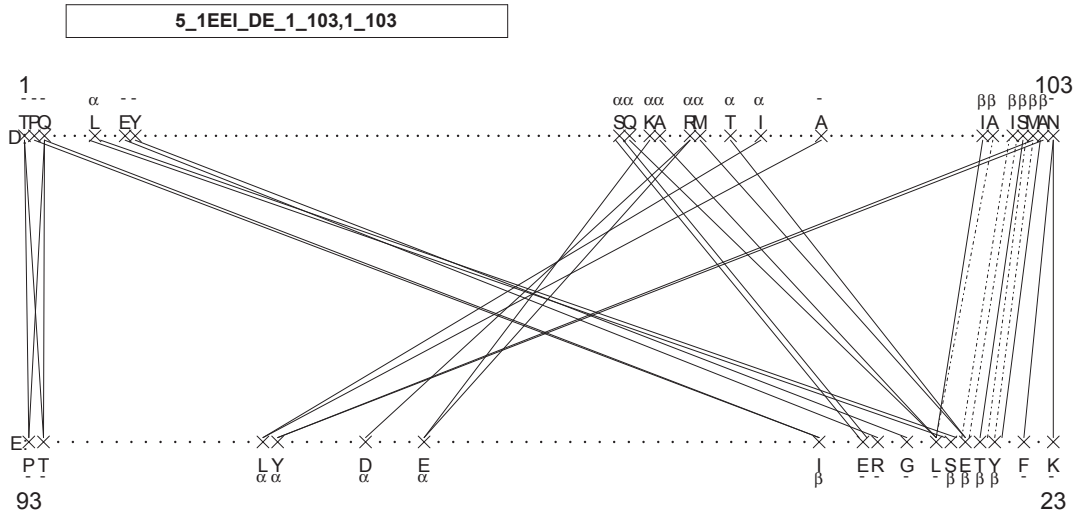


Figure 3.4 : Exemple de graphe Gemini : Graphe présente les hots pots qui sont en interaction intermoléculaires dans la toxine du choléra.

Selon la figure 3.4, les deux segments qui forment une région sont représentés par deux lignes horizontales parallèles systématiquement positionnées à une distance fixe, les interactions impliquant des atomes du squelette des acides aminés sont indiquées par des traits en pointillé alors que les interactions impliquant des atomes de la chaîne latérale sont indiquées par des traits continus.

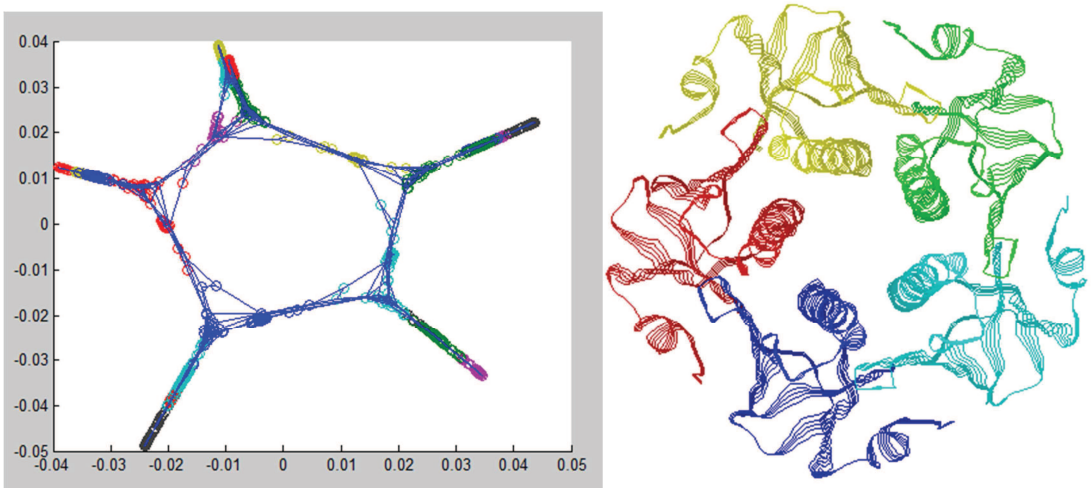
3.2.2 Spectral-pro

Cet environnement logiciel a été développé par Kavé SALAMATIAN et j'y ai aussi contribué sur certaines fonctionnalités. Spectral-pro a été initialement développé sous Matlab, mais récemment une version indépendante sous Python a été développée. Tout comme Gemini, Spectral-pro utilise les coordonnées cartésiennes des atomes d'une protéine obtenues à partir de la PDB, disponible dans la banque de données RSCB PDB (http://www.rcsb.org/pdb/home/home.do).

Spectral-pro construit un graphe où les nœuds sont les atomes de la protéine et chaque atome est connecté au plus à k autres atomes situés sur d'autres chaînes et pour des distances inférieures à 5 Å. Cette mesure de similarité est motivée par le fait qu'aucune interaction chimique faible n'est possible au-delà de 5Å. La construction du graphe aboutit à plusieurs sous graphes connexes (e.g. interface entre deux chaines dans un pentamère) impliquant plusieurs chaînes qui sont distribués symétriquement si l'oligomère est symétrique. Chacun de ces sous-graphes est représentatif d'une partie de l'interface. Spectral-pro applique une analyse spectrale sur chacune des composantes connexes individuellement. L'analyse spectrale permet ensuite de regrouper les nœuds à l'intérieure d'une composante connexe en groupes dans lesquels les nœuds sont fortement interconnectés ensembles et faiblement connectés avec des nœuds extérieurs au groupe.

Spectral-pro utilise le graphe entre les atomes de la protéine pour générer plusieurs autres graphes. Un premier graphe, que nous appellerons *unweighted* (sans pondération), connecte deux acides aminés si au moins un atome de chaque acide aminé est connecté dans le graphe initial. Un second graphe, qui est appelé *weighted* (pondéré), ajoute au graphe précédent un poids sur chaque lien représentant le nombre de liens exact entre atomes de chaque acide aminé. L'analyse spectrale est reproduite sur chacun de ces graphes permettant de regrouper les acides aminés d'une protéine en groupe d'acides fortement interconnectés à l'intérieur du groupe et faiblement connectés avec des acides aminés extérieurs au groupe.

Spectral-Pro génère plusieurs graphiques permettant de représenter les regroupements obtenus. Un graphique peut représenter les graphes dans leurs coordonnées spectrales (Figure 3.5).



(A) **(B)**

Figure 3.5 : A : Graphe unweighted de la toxine du choléra, les couleurs présentent les acides aminés d'un même groupe après analyse spectrale. B : Représentation de la structure atomique de la toxine du choléra (Rasmol), chaque couleur correspond à une chaîne. On peut noter que l'analyse spectrale met en évidence plusieurs groupes d'acides aminés constituant l'interface entre deux chaines.

Les composantes connexes (groupe après analyses spectrale) remarquées ne correspondent pas aux régions obtenues par Gemini. Cette observation montre qu'à l'intérieur d'une même région d'interface, il existe plusieurs types de connectivité.

Une autre possibilité est de zoomer sur les acides aminés dans chaque composante connexe (Figure 3.6).

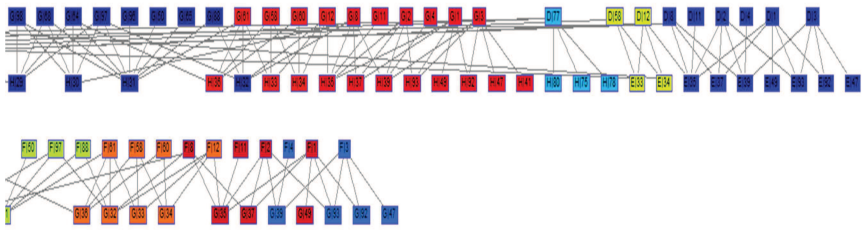


Figure 3.6 : une partie du graphe représente les liens entre les acides aminés. Les acides aminés de la même chaîne sont indiqués par couleur.

On peut aussi choisir une représentation graphique à partir de la structure atomique de la protéine (Figure 3.7). Ici la représentation graphique se fonde sur le logiciel de visualisation de protéine Jmol (ou Pymol).

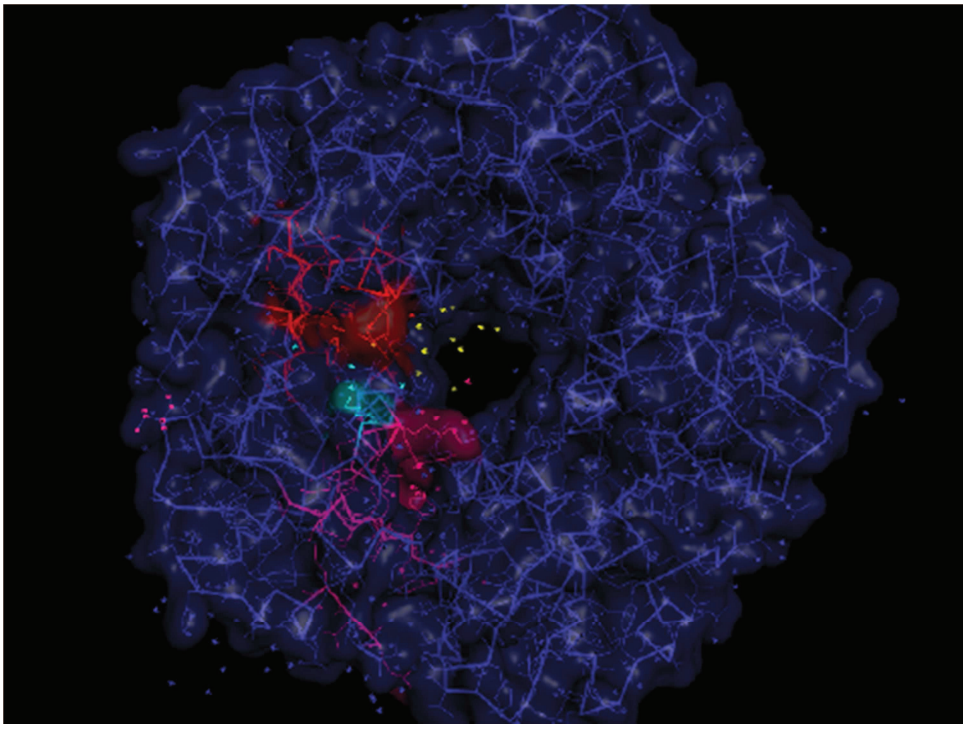


Figure 3.7 : figure représentant une interface dans la structure 3D de la toxine du choléra. Les zones colorées indiquent les interfaces de la toxine du choléra.

En plus de ces représentations graphiques, Spectral-Pro exporte les différentes matrices d'adjacence des graphes sont des formes exploitables par les logiciels d'analyse de graphes comme *Gephi* permettant de calculer des statistiques sur la structure du graphe.

3.2.3 Fold-X

3.2.3.1 Définition

Fold-X est un algorithme qui peut fournir une estimation rapide et quantitative de l'importance des interactions protéiques. Les différents termes d'énergie pris en compte dans Fold-X ont été pondérés à l'aide des données empiriques obtenues à partir d'expériences d'ingénierie des protéines [101].

Fold-X a été développé par le consortium Fold-X, centré autour du laboratoire de Luis Serrano au Laboratoire de Biologie Moléculaire Européen (EMBL) à Heidelberg.

La fonction d'énergie utilise un minimum de ressources de calcul et peut donc facilement être utilisées dans les algorithmes de conception de protéines, dans le domaine de la structure des protéines et là où l'on a besoin de calculs rapides et précis. Fold-X est

disponible via une interface web et peut être téléchargé pour une installation locale (`http://foldx.crg.es/about.jsp`).

Fold-X emploie un champ de force empirique qui a été développé pour l'évaluation rapide de l'effet des mutations sur la stabilité, le repliement et la dynamique des protéines et des acides nucléiques. La fonctionnalité de base de Fold-X est le calcul de l'énergie libre d'une macromolécule sur la base d'images 3D à haute résolution de sa structure.

La fonction d'énergie Fold-X comprend des termes qui ont été jugés importants pour la stabilité des protéines.

$$\Delta G = W_{vdw} * \Delta G_{vdw} + W_{solvH} * \Delta G_{solvH} + W_{solvP} * \Delta G_{solvP} + \Delta G_{wb} + \Delta G_{hbond} + \Delta G_{el} + \Delta G_{Kon} + W_{mc} * T * \Delta S_{mc} + W_{sc} * T * \Delta S_{sc}$$

Avec :

- ΔG_{vdw} : la somme des contributions de Van Der Waals des atomes par rapport à la même interaction avec le solvant. ΔG_{vdw} est calculé de façon similaire à la désolvatation, mais en prenant désormais en compte les énergies de transfert expérimentales de l'eau à la vapeur.
- ΔG_{solvH} et ΔG_{solvP} : sont les différences dans l'énergie de solvatation pour les groupes apolaires et polaires, respectivement, lorsque ceux-ci changent de l'état déplié à l'état replié. L'interaction avec le solvant est traitée en deux étapes: d'abord, le solvant en vrac est traité comme un terme de désolvatation qui est continuellement mise à l'échelle à l'enterrement d'un atome et séparé en contributions des groupes hydrophobes (ΔGsolvH) et polaires (ΔGsolvP).
- ΔG_{hbond} : est la différence d'énergie libre entre la formation d'une liaison hydrogène intra-moléculaire par rapport à la formation de liaisons hydrogène inter-moléculaires (avec solvant).
- ΔG_{wb} : est l'énergie de stabilisation fournie par une molécule d'eau présentant plus d'une liaison hydrogène de la protéine (ponts d'eau); les molécules d'eau qui ont une interaction avec les groupes persistante de la protéine, c'est à dire font plus de deux liaisons hydrogène avec la protéine, sont calculés de façon explicite dans l'expression ΔG_{wb}
- ΔG_{el} : est la contribution électrostatique de groupes chargés.

- ΔS_{mc} : est le coût entropique de fixation de la chaîne principale à l'état plié, ce terme dépend de la tendance intrinsèque d'un acide aminé particulier à adopter certains angles dièdres.
- ΔS_{sc} : est le coût entropique de la fixation d'une chaîne latérale dans une conformation particulière, est obtenue par mise à l'échelle d'un ensemble des paramètres d'entropie calculés par Abagayan.

Les termes $\mathbf{W_{vdw}}$, $\mathbf{W_{solvH}}$, $\mathbf{W_{solvP}}$, $\mathbf{W_{mc}}$ et $\mathbf{W_{sc}}$ correspondent aux facteurs de pondération appliqués aux termes d'énergie brute.

La plus grande précision dans les prédictions Fold-X est obtenue lorsque la différence d'énergie peut être calculée entre deux structures bien définies, par exemple entre la native et un mutant, ou entre les formes liées et non liées d'un complexe de protéines (pour déterminer la connexion de liaison d'énergie). La différence dans les énergies libres calculées ($\Delta\Delta G$) entre l'état final (le mutant) et l'état initial (la protéine native) est bien corrélée avec la variation observée expérimentalement dans la stabilité. D'autre part, l'énergie libre de pliage est calculée à partir de la différence de Gibbs d'énergie libre entre la structure 3D détaillée trouvé dans le fichier PDB et un état de référence déplié hypothétique pour lequel aucun détail structurel n'est disponible.

Fold-X peut muter un ou plusieurs acides aminés d'une structure donnée en le remplaçant par un ou les 20 acides aminés naturels, les versions phosphorylès de Ser, Thr et Tyr, la version sulfate, méthyle Lys, hydroxyle, Proline et la norme de 4 bases d'ADN.

3.2.3.2 Les opérations de Fold-X

Fold-X vise à décrire les contributions énergétiques à la stabilité des protéines en termes empiriques simples qui permettent une interprétation aisée par des non-spécialistes. Il est donc adapté à des tâches de bioinformatique structurelle à haut débit. Ici, on l'emploiera pour calculer l'énergie de la stabilité d'une protéine, l'énergie des interactions à l'interface d'un oligomère (ex. le pentamère de la toxine du choléra) et dans des versions de type sauvage ou mutées (Figure 3.8). L'exécution du programme nécessite de trois fichiers principaux [102]:

- Fichier PDB : est un fichier texte qui décrit la position des atomes des protéines dans l'espace. Fold-X calcule la position de tous les protons liés aux atomes de la protéine en utilisant un ensemble de coordonnées canoniques pour chaque type d'acide aminé.
- Fichier Run.txt : est une combinaison de commandes et d'options et permet de construire des scripts pour Fold-X.
- Fichier list.txt : est un fichier où on indique les noms des fichiers PDB à traiter.

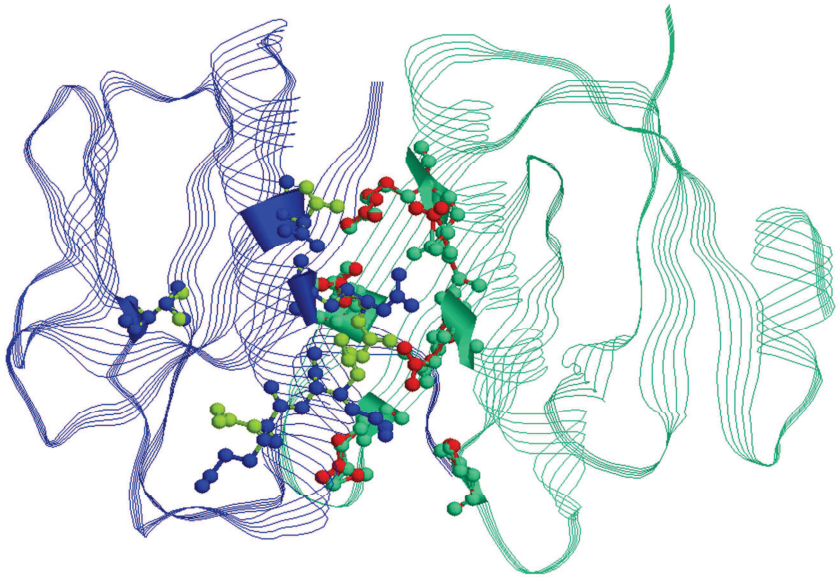


Figure 3.8 : Alignement de structures de deux chaines du pentamère de la sous unité B de la toxine du choléra. Chaînes bleu et verte version native, chaîne jaune et rouge version mutée (Mutation K69N). Alignement produit par le programme Calpha matching.

La réparation de PDB

Cette opération fait une optimisation rapide de la structure. En particulier, elle traite toutes les chaînes latérales pour éliminer les « affrontements » de Van Der Waals, elle identifie les résidus qui ont des angles de torsion non favorables, et les repositionnent. De cette façon, les angles ou distances non-standard dans la structure sont évités afin de ne pas donner des valeurs d'énergie irréalistes.

Après réparation du fichier PDB, il est conseillé à l'utilisateur de comparer la structure d'origine avec celle réparée afin de vérifier les changements et les corriger si besoin, par exemple si un résidu qui ne doit pas bouger a été déplacé. Dans un tel cas, le fichier « run.txt » doit être modifié (**Annexe 2**). Le résidu peut être fixé avant la réparation en utilisant un autre fichier run.txt (**Annexe 3 (a)**).

La stabilité

Une fois la structure PDB réparée, l'utilisateur peut commencer à travailler avec elle. Il peut par exemple lancer la commande «Stability » pour effectuer le calcul de l'énergie de stabilité de la molécule, en changeant simplement le nom « PDB » dans le fichier list.txt par le nom du fichier « RepairPDB_1EEI.pdb »**Annexe 3 (b)**).

Interaction entre les interfaces

L'analyse des énergies d'interaction nécessite un complexe protéique. Dans le cas étudié, comme les chaînes du pentamère de la toxine du choléra sont identiques, seules deux chaînes adjacentes sont considérées pour les calculs des énergies d'interactions (c'est-à-dire l'énergie de l'interface). Pour ce calcul, un autre fichier « run.txt » doit être crée (**Annexe 3(c)**).

Mutation des PDBs

Lorsqu'on essaye d'améliorer la stabilité d'une protéine, il n'est pas facile de décider quelles positions mutées car on ne peut pas anticiper les effets des mutations sur la stabilite. Une possibilité est de demander à Fold-X d'analyser certaines positions dans la séquence en exécutant soit la commande <PositionScan> soit <BuildModel>. De cette façon, Fold-X va muter les résidus dans les positions choisies vers un autre type d'acides aminés (tous les cas sont essayés).

La fonction mutation peut être faite sur une ou plusieurs mutations dans une ou plusieurs positions dans la même simulation. Il existe trois façons pour réaliser des mutations :

- ***Mutation individuelle :*** cette opération nécessite de choisir la position des groupes de résidus c'est-à-dire dans le fichier run.txt on note le résidu souhaite muté. Un run.txt typique ressemblera (**Annexe 3 (d)**) aux résidus et positions mutées.
- ***Multiples mutations :*** en utilisant un fichier mutant ou on montre toute la séquence souhaité mutée. Cette opération nécessite un fichier séparé du fichier run.txt (**Annexe 3 (e)**).
- ***Multiples mutations :*** en utilisant un fichier individuel_list ou on note la liste des mutations souhaitées. Dans ce cas, le fichier individual_list.txt est séparé du fichier

run.txt (**Annexe 3 (f)**) et doit ressembler chaque résidu accompagné de sa mutation choisie.

Les fichiers de sortie seront des fichiers PDB avec un nom pris par le résidu et la position mutée.

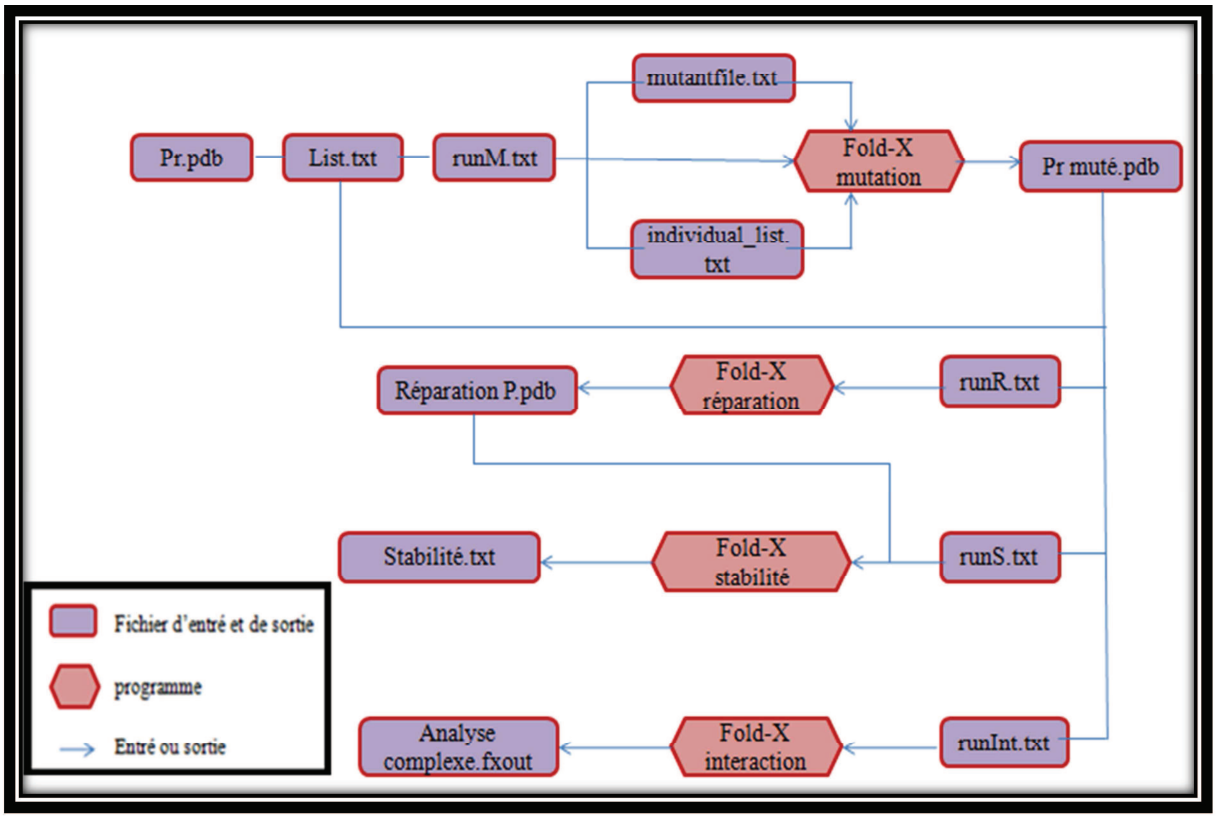


Figure 3.9 : organigramme général pour lancer les simulations dans Fold-X.

Chapitre 4: Communication entre réseau d'acides aminés intramoléculaires et réseau d'acides aminés intermoléculaires

L'implication de la déprotonation des histidines dans l'assemblage de la toxine du choléra ($CtxB_5$), résultat expérimental, mène à se poser la question : comment les histidines régulent-elles l'interface protéique de la $CtxB_5$? Ma thèse commence par cette question. J'ai suivi l'étude du rôle du résidu histidine 94 en utilisant des outils informatiques et en particulier une approche réseau.

Pour répondre à la question, il faut d'abord connaitre la position des histidines dans le pentamère $CtxB_5$. La première étape était la visualisation des histidines sur la structure atomique de $CtxB_5$. J'ai pu noter qu'aucune histidine n'étaient localisées dans les interfaces. Ce résultat a été confirmé par l'utilisation des outils informatiques développés par notre équipe, Gemini et Spectral-Pro, qui permettent d'identifier tous les résidus aux interfaces à partir de la structure atomique. Donc comment les histidines régulent l'interface sans être parmi les acides aminés formant l'interface ? L'hypothèse était que les histidines régulent l'interface indirectement c'est-à-dire l'existence d'une communication indirecte entre les histidines et l'interface via des résidus intermédiaires (les résidus ayant des interactions intramoléculaires avec les histidines et impliqués dans l'interface) proche des histidines. En regardant la structure atomique, l'His 94 est trop loin de l'interface (la distance est supérieure à 5Å), mais parmi les résidus en contact chimique avec l'His 94 certains font partis de l'interface. Cela suggère une communication via une distance géodésique c'est-à-dire le chemin de communication entre l'His 94 et l'interface via le résidu 49 par exemple. Donc l'His 94 influence sur le résidu 49 qui influence sur le résidu 1. L'hypothèse repose sur la possibilité d'avoir des perturbations locales sur une histidine se répercutant au-delà de ses voisins chimiquement proches. J'ai considéré tous les chemins possibles entre l'His 94 et les résidus de l'interface et j'ai cherché à montrer s'il y avait ou non une communication/ influence entre ces résidus qui communiquent l'His 94 et l'interface. La mutation individuelle, double et triple a pour objectif de tester si le phénomène d'additivité existe ou non. Un chemin de communication existe si l'effet non additif (l'énergie de la triple mutation est différente à celle de la double) sur l'énergie d'interaction des mutations des résidus le long du chemin existe. Les résultats ont montré qu'il existe une communication entre His 94 et l'interface.

Pour conforter les résultats obtenus, la comparaison entre le pentamère de la toxine thermolabile B (LTB_5) et le pentamère de la toxine cholérique B ($CtxB_5$) nous a permis de

mettre en évidence différents chemins de communication entre les résidus des interfaces et les résidus des chaînes individuelles hors interfaces dans les deux toxines. Le but de cette comparaison est de savoir si la position du résidu 94 est stratégique et lui confère ce rôle de régulateur ou si c'est le type de résidu à cette position. La même procédure a été testée sur la toxine LTB_5, les résultats ont montré que les influences ne sont pas seulement le fait de la position d'un résidu mais aussi de son type à une position donnée, et de son environnement.

Les effets d'influence entre les résidus suivent un mécanisme en cascade se produisant de résidus voisins en résidus voisins.

Chapitre 4 : Communication entre réseau d'acides aminés intramoléculaires et réseau d'acides aminés intermoléculaires

Mounia Achoch[1], Giovanni Feverati[2], Kave Salamatian[1], Rodrigo Dorantes-Gilardi[3], Laurent Vuillon[3] and Claire Lesieur[4].

[1]Laboratoire d'informatique système, traitement de l'information et de la connaissance (LISTIC), Université de Savoie, Annecy le Vieux, France;

[2]Federation de recherche Fr3405, Modelisation, Simulations, Interactions Fondamentales, Annecy-le-Vieux, France

[3]Laboratoire de mathématiques (LAMA UMR 5127), Université Savoie Mont Blanc, CNRS, Le Bourget du Lac, France;

[4]CNRS—ENS-Lyon-UCBL, Laboratoire AMPERE, Lyon, France

Résumé

L'assemblage protéique est un mécanisme de combinaison de deux ou plusieurs chaînes protéiques, ce mécanisme est souvent utilisé par des organismes vivants pour déclencher une activité biologique. Le pentamère de la sous unité B de la toxine du choléra ($CtxB_5$), qui appartient à la famille des toxines AB_5, est présenté ici comme modèle d'étude de l'assemblage des protéines.

Des résultats expérimentaux ont montré l'importance des résidus d'histidine dans l'assemblage de $CtxB_5$. Il s'agit ici de comprendre comment les résidus d'histidine régulent l'assemblage alors qu'ils sont localisés à l'extérieur de l'interface de la toxine. Le rôle du résidu histidine 94 est exploré. L'étude de cette histidine est effectuée en utilisant des outils informatiques et en particulier une approche réseau (voir méthodes). La comparaison entre le pentamère de la toxine thermolabile B (LTB_5) et le pentamère de la toxine cholérique B ($CtxB_5$) nous a permis de mettre en évidence différents chemins de communication entre les résidus des interfaces et les résidus des chaînes individuelles hors interfaces dans les deux toxines. Ces résultats ouvrent des pistes pour comprendre pourquoi ces deux toxines suivent différents mécanismes d'assemblage.

4.1 Introduction

La modélisation informatique et mathématique de phénomènes physiques et biologiques complexes est un véritable défi posé aux informaticiens et mathématiciens. Devant l'arrivée en masse de données biologiques, il devient important de fournir des modèles permettant d'exploiter ces données afin d'aider à comprendre les phénomènes qui rentrent en jeu. Dans cet article, nous exposons un problème biologique et proposons une modélisation à l'aide de systèmes informatiques (notions de réseaux).

La fonction d'une grande majorité des protéines dépend de leur capacité à l'auto-assemblage, soit transitoire ou permanent, en oligomères [103, 104]. Comprendre les mécanismes d'assemblage des protéines est particulièrement important en raison de l'implication d'oligomères dans de nombreuses pathologies, de l'infection bactérienne aux maladies conformationnelles (ex. la maladie d'Alzheimer ou la maladie de Parkinson) [105, 106].

La toxine du choléra est le facteur de virulence le plus importante produit par *Vibrio cholerae*. Cette toxine est composée de deux sous unités A et B. La sous-unité B est un pentamère principalement impliqué dans le transport de la sous-unité A vers sa cible dans des cellules [22, 107]. La sous-unité B peut être produite par les bactéries, en l'absence ou en présence de la sous-unité A, ce qui signifie que le pentamère est une entité structurellement et fonctionnellement indépendante [108].

L'entéro-toxine thermolabile (LTB_5) comme la toxine du choléra ($CtxB_5$) est un complexe hétéro-héxamèrique [109-111]. Les deux sous-unités B partagent une identité de séquence de 94% [23].

Les structures cristallographiques de LTB_5 et $CtxB_5$ sont quasi superposables, de formes circulaires et chaque sous-unité B interagit largement avec ses sous-unités adjacentes [112-114]. Par conséquent, les pentamères sont très stables, et ne se dissocient qu'à des pH inférieur à 2,0 [115, 116]. La différence peut être localisé à quatre et à deux histidines qui sont répartis le long de la séquence des protéines $CtxB_5$ et LTB respectivement et peuvent agir ensemble.

Le réassemblage du $CtxB_5$ *in vitro* est inhibé entre les pH 5,0 et 8,0, avec un pKa à environ 6,0 [108]. Le réassemblage de LTB_5 est inhibe dans la même gamme de pH mais a un pKa autour de 7.0 [117]. La différence est imputée au nombre d'histidine présentés dans les deux toxines, $CtxB_5$ a quatre histidines (His 13, His 18, His 57 et His 94) alors que LTB_5 n'en a que deux (His 13 et His 57). Les histidines 18 et 94 sont positionnées en amont des deux brins β composant l'interface principale des toxines (Figure 4.1).

Les résultats obtenus expérimentalement combinés à des simulations confortent l'implication de la déprotonation des histidines dans l'assemblage de $CtxB_5$ [108]. L'assemblage de LTB_5 implique la déprotonation du N-terminal de la toxine, le résidu alanine en position 1 est un des hots spots de l'interface de LTB_5, en bon accord avec son implication dans l'assemblage [23].

Par contre aucune histidine n'est localisée dans l'interface de $CtxB_5$, ce qui rend plus difficile leur implication dans l'assemblage et suggère un mécanisme de régulation indirect. Le but de ce travail est de trouver le lien entre les résidus d'histidines et l'interface qui permette d'expliquer leur implication dans l'assemblage de la toxine du choléra.

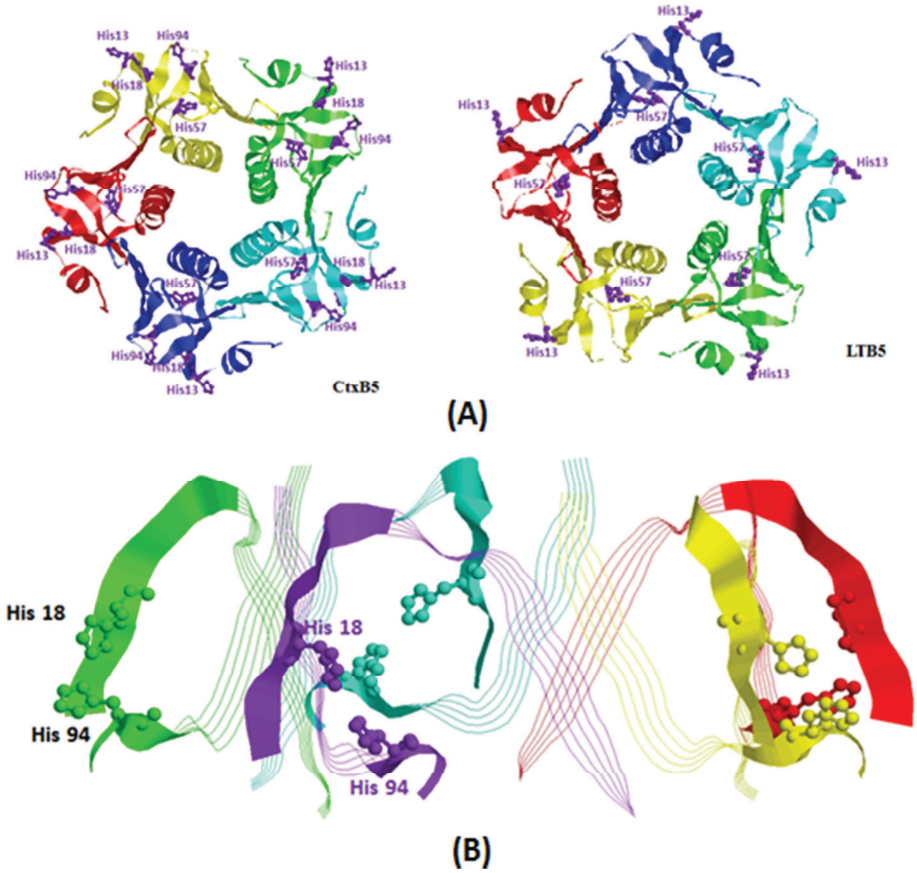


Figure 4.1 : A. Positions des histidines dans les deux toxines $CtxB_5$ et LTB_5. B. Zoom sur les histidines 18 et 94, Les brins en représentation "strands" correspondent aux deux brins de l'interface principale, brin β 23-31 et brin β 96-103.

4.2 Méthode

Fold-X : http://foldx.crg.es/ (détails dans le chapitre 3, méthodologies)

Génération des structures de mutants

Les structures mutantes ont été générées en utilisant la fonction de position <PositionScan> ou <BuildModel> de Fold-X. Au cours de cette procédure de conception, Fold-X teste différents rotamères et permet aux atomes des chaînes latérales voisines de se déplacer. Le programme introduit d'abord une mutation en alanine, puis mute dans le résidu désiré (tout en déplaçant les résidus voisins).

Optimisation des modèles en utilisant la fonction de réparation Fold-X

Les structures 3D ont été prises la Protein Data Bank (PDB 1EEI), et soumises à une procédure d'optimisation en utilisant la fonction de réparation de Fold-X. Au cours de cette procédure, Fold-X identifie les résidus qui ont angles faible de torsion, exposition de Van Der

Waals, ou affrontements énergies totales. Fold-X fonctionne comme suit: premièrement, il mute la position sélectionnée en un résidu alanine et annote les énergies des chaînes latérales des résidus voisins. Puis il mute l'alanine en l'acide aminé sélectionné, et recalcule les énergies des chaînes latérale des mêmes résidus voisins. Ces résidus qui présentent une différence d'énergie sont alors mutés pour eux-mêmes, afin d'identifier le rotamère d'énergie la plus favorable. Cette procédure contient une fonction supplémentaire, où Fold-X élimine rapidement les petits affrontements locaux, et enregistre le temps de calcul.

Calculs de l'énergie

Les calculs de l'énergie des protéines mutantes ont été effectués avec la fonction d'énergie de Fold-X qui comprend les termes considérés comme importants pour la stabilité des protéines.

Les valeurs des énergies obtenues par Fold-X sont ensuite converties en valeurs plus réalistes grâce à l'utilisation d'une fonction de normalisation obtenue en ajustant les données expérimentales et calculées.

Gemini (Détails dans le chapitre 3, méthodologies)

Ce programme crée une représentation graphique des régions d'interface et de leurs interactions entre acides aminés dans le style de la théorie des graphes. Ici, les sommets du graphe sont les acides aminés; ceux qui sont sélectionnées comme participants à une liaison chimique faible sont symbolisés par des croix "X" (Hots spots) (voir chapitre 3, méthodologies), les autres sont symbolisés par des points "." (mentionné ici la figure du chapitre 3 en exemple de Gemini graph). Pour modéliser une interface, Gemini tient compte de tous les acides aminés des interfaces et pas seulement ceux impliqués dans des interactions chimiques.

Spectral-Pro

Cet algorithme développé aussi sur la théorie des graphes, se base sur les matrices d'adjacences et les propriétés spectrales des réseaux pour modéliser les interfaces sous la

forme d'un réseau d'acides aminés en interaction. Les propriétés des interfaces sont inférées les propriétés de leurs réseaux.

4.3 Résultats

Le but de ce travail est d'expliquer comment les histidines régulent l'assemblage protéique de $CtxB_5$ tout en étant localisées en dehors de l'interface. Comme décrit dans le chapitre 1 de l'introduction, l'assemblage d'une protéine (ex. $CtxB_5$) combine des réactions de repliement et d'association. Ainsi, pour avoir une image complète du mécanisme de réassemblage, en plus d'étudier les interactions intermoléculaires des acides aminés (formation de l'interface), il est nécessaire d'étudier les interactions intramoléculaires des acides aminés et d'appréhender comment des deux types d'interactions sont coordonnés [118, 119]. Pour répondre à cette question, j'ai utilisé des approches et outils dérivés de l'informatique tout en me reposant sur des données expérimentales de l'assemblage de $CtxB_5$.

Au laboratoire, il a été montré expérimentalement et par des simulations que l'assemblage de la toxine du choléra $CtxB_5$ (1EEI) était inhibé à pH acide avec un pKa autour de 6.0 suggérant l'implication des résidus histidines dans la régulation de l'assemblage. Cependant aucune histidine n'est localisée dans l'interface (Figure 4.2A) et le programme Gemini ne les identifient pas comme hots spots (Figure 4.2B). Afin de vérifier que les histidines n'avaient pas d'effet direct sur l'interface, les résidus histidines ont été mutés en asparagine et l'effet des mutations a été mesuré par des calculs d'énergie d'interaction de l'interface après et avant mutation (Tableau 4.1).

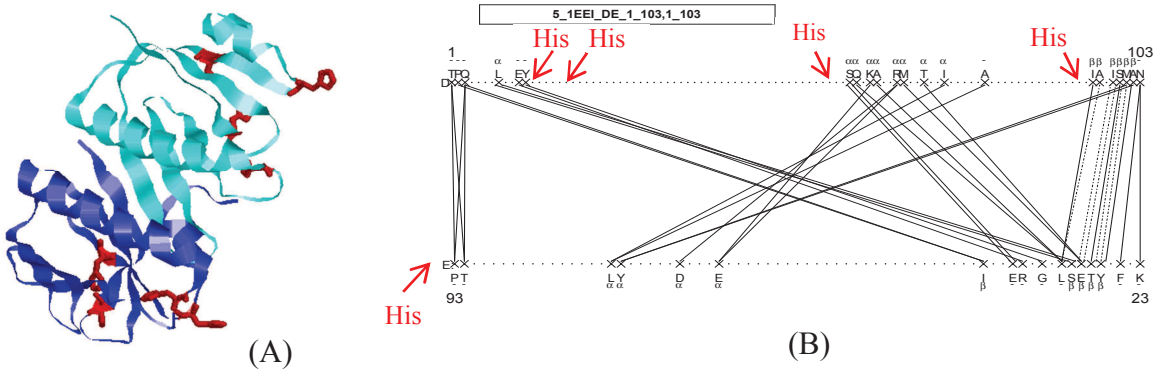


Figure 4.2 : Histidines dans $CtxB_5$. A. Représentation par Rasmol de deux chaînes sur les cinq du pentamère, désignée par des couleurs différentes. Les histidines sont colorées en rouge. B. Graphe Gemini de l'interface de $CtxB_5$. Les histidines ne sont pas identifiées comme des hots spots et leurs positions sont indiqués par des flèches rouges.

Mutant	Energie d'interaction (Kcal/mol)
WT	**-13,75**
H13N	**-13,75**
H18N	**-13,76**
H57N	**-13,67**
H94N	**-13,75**

Tableau 4.1 : les valeurs d'énergie d'interaction des quatre histidines mutées

Le programme Fold-X est utilisé pour générer les mutations et effectuer les calculs des énergies d'interaction. La mutation individuelle des histidines n'a aucun effet sur l'énergie d'interaction. Les mutations doubles, triples et quadruples des histidines ont aussi été testées et aucune ne mène à une modification de l'énergie d'interaction (Tableau 4.2).

Mutant	Energie d'interaction (Kcal/mol)
H13N	-13,75
H18N	-13,76
H57N	-13,67
H94N	-13,75
H13N+H18N	-13,76
H57N+H94N	-13,59
H13N+H57N	-13,47
H13N+H94N	-13,75
H18N+H57N	-13,88
H18N+H94N	-13,73
H13N+H18N+H57N	-13,5
H13N+H57N+H94N	-12,97
H18N+H57N+H94N	-13,58
H13N+H18N+H57N+H94N	-13,51

Tableau 4.2 : les énergies d'interaction des mutations doubles, triples et quadruples des histidines.

Puisqu'aucune histidine n'est localisée dans l'interface et que leur mutation individuelle ou combinée n'a pas d'effet direct sur l'interface, nous avons émis l'hypothèse que les histidines régulaient l'assemblage indirectement en intervenant sur des acides aminés impliqués dans l'interface. L'objectif est de tester s'il existe une communication entre les histidines et l'interface via des résidus « intermédiaires » proche des histidines, c'est à dire ayant des interactions intramoléculaires avec les histidines et impliqués dans l'interface. Cette hypothèse repose sur la possibilité d'avoir des perturbations locales sur une histidine se

répercutant au-delà de ses voisins chimiquement proches. Un tel phénomène est classique en biologie, il s'agit de l'allostérie. La notion de réseau est très pertinente pour comprendre les mécanismes de communication d'une échelle locale a une échelle globale [120].

L'histidine 94 a 50% de ses voisins intramoléculaires chimiquement proches qui sont des hots spots (Tableau 4.3). De plus ses hots spots sont impliqués dans deux régions d'interface différentes localisés sur deux chaînes différentes. Considérant l'histidine 94 de la chaîne E comme exemple, ses voisins 47, 49, 92 et 93 interagissent intermoléculairement avec les résidus 1 à 3 sur la chaîne D alors que ses voisins 88 et 96 intéragissent intermoléculairement avec les résidus 23 à 31sur la chaîne F (Figure 4.3). Ce résidu est de ce fait en position stratégique puisqu'il est en interaction avec des résidus du réseau intermoléculaire sans en faire partie lui-même et permet de connecter indirectement trois chaînes entre elles. L'histidine 13 a 25 % de ses voisins intramoléculaires qui sont des hots spots (résidus 11, 12 et 88), ses voisins intéragissent intermoléculairtement avec les résidus de 31 à 35. L'histidine 57 a 33 % de ses voisins intramoléculaires qui sont des hots spots (résidus 61, 58, 53 et 65), ses voisins intéragissent intermoléculairement avec les résidus de 31 à 36 de la chaîne F et le résidu 63 de la chaîne D. L'histidine 18 n'a aucun hot spot en voisin intramoléculaire.

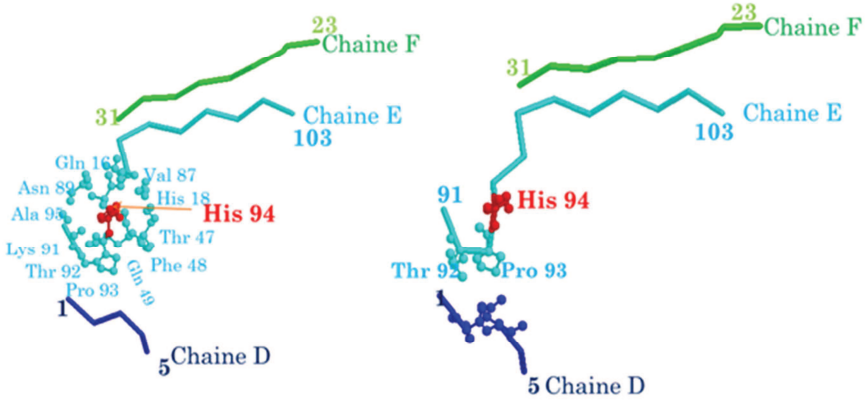


Figure 4.3 : Voisinage de l'histidine 94 (Rasmol). L'histidine 94 de la chaîne E (couleur turquoise) est indiquée en rouge, ces voisins intramoléculaires localisés à une distance de 5 Å sont indiqués en représentation boules et bâtons. Leurs voisins intermoléculaires sont indiqués en représentation squelette en bleu pour l'interface sur la chaîne D et en vert pour l'interface sur la chaîne F.

Le travail se focalise sur le cas du résidu histidine 94. L'approche réseau utilisé pour ce travail est basé sur l'établissement de chemins de communication entre nœud d'un réseau. Il s'agit ici de considérer tous les chemins existant entre l'histidine 94, ses voisins

« intermédiaires » et leurs voisins intermoléculaires (**Annexe 4**). J'ai alors testé tous ces chemins de communication en mutant en asparagine tous les résidus le long du chemin, individuellement puis en combinaison et en calculant avec Fold-X les effets sur l'énergie d'interaction. En gros, un effet non additif sur l'énergie d'interaction des mutations des résidus le long du chemin permet de suggérer un chemin de communication. Les résultats ont été analysés en considérant aussi les effets des mutations sur les réseaux de l'interface, produit par Spectral-Pro et Fold-X.

	Intramoléculaire	**Intermoléculaire**
His 94	Gln 16	***
	His 18	***
	Thr 47	Gln 3
	Phe 48	***
	Gln 49	Thr 1
	Val 87	***
	Trp 88	Leu 31, Ala 32
	Asn 89	***
	Lys 91	***
	Thr 92	Thr 1, Gln 3
	Pro 93	Thr 1, Pro 2, Gln 3
	Ile 96	Leu 31

Tableau 4.3 : les acides aminés en interaction intramoléculaire avec l'histidine 94 et leurs interactions en intermoléculaire de chaque acide aminé dans la toxine $CtxB_5$. (***) : pas de liaison avec l'interface

Deux exemples sont donnés en Tableau 4.4 pour illustrer d'abord un chemin de communication identifié par des effets non additifs des mutations combinées et un autre chemin sans communication, défini comme pas d'effet non additif des mutations des résidus le long du chemin.

$CtxB_5$	
Mutants	**Energie d'interaction (Kcal/mol)**
WT	**-10,9**
H94N	-10,9
Q49N	-11,3
H94N+Q49N	-11,4
T1N	-11,0
H94N+T1N	-11,1
Q49N+T1N	-13,1
H94N+Q49N+T1N	**-11,7**

$CtxB_5$	
Mutants	**Energie d'interaction (Kcal/mol)**
WT	**-10,9**
H94N	-10,9
I96N	-10,8
H94N+I96N	-10,9
L31N	-8,4
H94N+L31N	-8,8
I96N+L31N	-8,8
H94N+I96N+L31N	**-8,9**

Tableaux 4.4 : les mutations individuelles, doubles et triples de deux chemins de communications dans la toxine du choléra ($CtxB_5$).

La notion d'additivité : c'est une propriété des grandeurs mesurables. C'est une approche utilisée pour l'évaluation des mélanges. La notion d'additivité est utilisée dans le cas de mélanges relativement simples comprenant au plus une douzaine de composés. Le concept d'interaction comprend tous les cas où les effets d'un mélange est différent de l'additivité (Casse et al. 1998).

L'additivité est présenté en deux parties :

- Additivité intra (codominance) : les deux mutations sont indépendantes l'une de l'autre

$$AA \rightarrow x_1$$
$$aa \rightarrow x_2$$
$$Aa \rightarrow \frac{x_1 + x_2}{2}$$

- Additivité inter (épistasie) : les mutations s'influencent les unes les autres.

$$A \rightarrow \alpha_1 \qquad AB \rightarrow \alpha_1 + \beta_1$$
$$a \rightarrow \alpha_2 \qquad Ab \rightarrow \alpha_1 + \beta_2$$
$$B \rightarrow \beta_1 \qquad aB \rightarrow \alpha_2 + \beta_1$$
$$b \rightarrow \beta_2 \qquad ab \rightarrow \alpha_2 + \beta_2$$

Quand les mutations multiples donnent une énergie égale à la somme des énergies des mutations individuelles cela signifie que les mutations sont additives, elles sont donc indépendantes les unes des autres et il n'y a pas de communication. Les mutations H94, T1, Q49 montrent que les trois mutations ont une énergie qui n'est pas égale à la somme des énergies pour les mutations individuelles. En gros comme H94N n'a pas d'effet et que la mutation double Q49N+T1N a une énergie de -13, 1 Kcal/mol, la triple mutation aurait une énergie de -13.1 Kcal/mol si les trois mutations étaient additives. Ce qui n'est pas le cas, donc ces mutations s'influencent entre elles, suggérant un chemin de communication entre les trois résidus. Le deuxième chemin qui implique les résidus H94, I96, et L31 présente une situation ou les mutations combinées ont des effets additifs suggérant qu'il n'y a pas de communication entre ses résidus. En effet, seule la mutation du résidu L31N modifie l'énergie d'interaction et les mutations combinées ont quasiment le même effet que la mutation seule. La communication entre l'histidine 94 et la région d'interface impliquant les résidus 23 à 31 ne semble donc pas impliquer le résidu I96. D'après les autres chemins possibles vers cette région d'interface, une autre possibilité implique le résidu W88 (**annexe 4**). Les mutations combinées permettent de montrer une non additivité suggérant un chemin de communication via le résidu W88 (Tableau 4.5).

$CtxB_5$	
Mutants	**Energie d'interaction (Kcal/mol)**
H94N	-11
W88N	-11
L31N	-8
H94N+L31N	-9
H94N+W88N	-11
W88N+L31N	-9
H94N+W88N+L31N	-11

Tableaux 4.5 : les mutations individuelles, doubles et triples des résidus H94, L31 et W88 dans le chemin de communication dans la toxine du choléra ($CtxB_5$). (***) : pas de liaison avec l'interface

Pour conclure, les résultats suggèrent qu'il existe des chemins de communication privilégiés qui permettent à l'histidine 94 de réguler la formation des deux régions d'interface (Figure 4.3).

La prochaîne question qui m'a intéressée est de savoir si la position du résidu 94 est stratégique et lui confère ce rôle de régulateur ou si c'est le type de résidu à cette position qui est en jeu. Pour répondre à cette question, nous pourrions tester la mutation du résidu 94 par les 18 autres acides aminés et refaire des mutations doubles et triples. Cette approche est possible mais fastidieuse. A la place, nous avons choisi de considérer le cas de la LTB_5 qui a un résidu asparagine en position 94. J'ai donc appliqué la même procédure sur LTB_5 pour déterminer s'il existait les mêmes chemins de communication que ceux identifiés dans $CtxB_5$ et distinguer si ces influences sont dues à la position du résidu ou au type de résidu à cette position. Les résidus voisins intramoléculaires du résidu Asn94 et leurs voisins intermoléculaires sont indiqués en Tableau 4.6.

	Intramoléculaire	**Intermoléculaire**
Asn 94	Tyr 18	***
	Thr 47	Gln 3
	Phe 48	***
	Gln 49	Ala 1
	Trp 88	Met 31, Ala 32
	Asn 89	***
	Asn 90	***
	Lys 91	***
	Thr 92	Ala 1, Gln 3
	Pro 93	Ala, Pro 2, Gln 3
	Ser 95	***
	Ile 96	Met 31

Tableau 4.6 : Résidus intramoléculaires de N94 et leur voisinage intermoléculaire dans la toxine LTB_5.

Le tableau 4.5 pour LTB_5 montre que les chemins entre voisins intramoléculaires du résidu N94 et leurs voisins intermoléculaires impliquant les résidus Q49, A1, I96, W88 et M31 existent aussi dans la LTB.

Seule la mutation A1N a un effet sur l'énergie d'interaction, effet qui n'est pas compensé en combinaison double (A1N+ Q49N ou A1N +N94H) ou triple (A1N+ Q49N+N94H) indiquant un effet additif des mutations et donc pas d'évidence d'un chemin de communication entre ces résidus. Ce résultat suggère donc que la position du résidu 94 n'explique pas à elle seule les influences et que le type d'acide aminé à cette position est aussi important. Cependant si la double mutation A1N + Q49N, qui est équivalente à la triple mutation dans $CtxB_5$, est considérée, on voit que l'effet de la mutation du résidu A1N n'est encore pas compensé, au contraire de ce qui est observé dans $CtxB_5$. Comme le résidu 94 est du même type dans ces deux situations, la différence d'influence est dû à un phénomène certainement plus complexe, impliquant peut être non seulement le type de résidu à la position 94 mais aussi les acides aminés en son voisinage. En effet, il faut noter que parmi les 19 résidus de compositions différentes dans les deux toxines, quatre sont dans cette région : Asn 94 (His 94), Met 31 (Leu 31), Tyr 18 (His 18), Ala 1 (Thr 1). De plus, le voisinage intramoléculaire du résidu 94 diffère aussi légèrement, de trois résidus sur douze. A la position 18 on observe une tyrosine pour LTB_5, mais une histidine pour $CtxB_5$; l'asparagine 90 et la serine 95 sont dans le voisinage du résidu asparagine 94 alors qu'à la place on observe les résidus glutamine 16 et valine 87 pour $CtxB_5$.

De la même manière, on ne retrouve pas le même résultat d'influence pour le chemin impliquant les résidus 94, 96 et 31 (Tableau 4.7). La double mutation (I96N + M31N) a une énergie de -6,3 Kcal/mol alors que la triple mutation (I96N + M31N + N94H) a une énergie de -8.4 Kcal/mol, indiquant un effet non additif des mutations et donc un potentiel chemin de communication entre ces résidus.

LTB_5	
Mutants	**Energie d'interaction (Kcal/mol)**
WT	-6,0
N94H	-6,0
Q49N	-5,5
Q49N+N94H	-6,0
A1N	-7,5
A1N+N94H	-7,3
Q49N+A1N	-7,1
Q49N+A1N+N94H	**-7,9**

LTB_5	
Mutants	**Energie d'interaction (Kcal/mol)**
WT	-6,0
N94H	-6,0
I96N	-5,2
N94H+I96N	-6,1
M31N	-3,9
M31N+N94H	-6,8
I96N+M31N	-6,3
N94H+I96N+M31N	**-8,4**

Tableaux 4.7 : les mutations individuelles, doubles et triples de deux chemins de communications dans la toxine labile (LTB_5).

Dans l'ensemble, les résultats montrent que les influences ne sont pas seulement le fait de la position d'un résidu mais aussi de son type à une position donnée, et de son environnement. En effet dans le deuxième chemin, le voisinage du résidu 94 est aussi diffèrent dans les deux toxines puisqu'on a une méthionine et une leucine dans LTB_5 et $CtxB_5$, respectivement. On commence à percevoir qu'il existe des effets d'influence entre les résidus suivant un mécanisme en cascade se produisant de résidus voisins en résidus voisins (Figure 4.4).

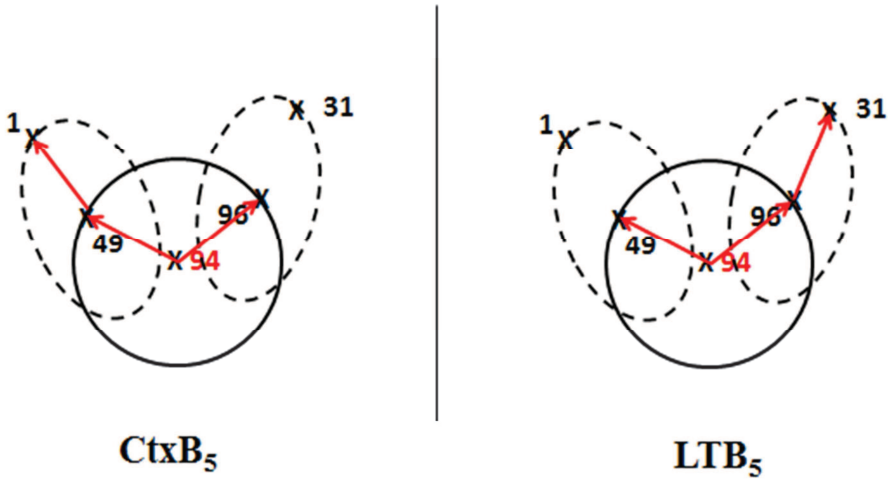


Figure 4.4 : représentation de la différence entre les deux toxines $CtxB_5$ et LTB_5 au niveau des chemins de communication. Le phénomène de l'influence en cascade est indiqué par les flèches rouges.

LTB_5 et $CtxB_5$ partagent une homologie de séquence de 94% (une identité de 82%) et des structures atomiques presque superposables mais néanmoins s'assemblent par deux

mécanismes différents (chapitre 1). La mise en évidence de chemins de communication entre acides aminés du réseau intramoléculaire et hots spots du réseau intermoléculaire différents liés à des acides aminés de compositions différentes offre une piste intéressante à explorer pour tenter de comprendre la différence de mécanisme d'assemblage entre les deux toxines. Ces résultats suggèrent aussi qu'il est fondamental de considérer non pas les acides aminés individuellement mais aussi leur voisinage pour appréhender l'étude des mécanismes de repliement et d'assemblage.

Chapitre 5: Mécanisme de communication en cascade : a-t-il aussi lieu dans le réseau intermoléculaire?

Le chapitre précédent comporte à étudier les chemins de communication entre le résidu d'His et les résidus des interfaces. Les résultats obtenus m'ont permis de tester si le mécanisme d'influence en cascade existe entre les résidus de l'interface, c'est à dire entre hots spots et pas entre les résidus intra-chaîne et les résidus de l'interface. Pour cela, j'ai cherché s'il existait une communication de pair à pair (communication entre acides aminés de proche en proche) entre acides aminés chimiquement voisins dans une région d'interface via un réseau d'interaction. Ce dernier est un ensemble des interactions entre les acides aminés des deux chaînes protéiques adjacentes. Mon prototype d'étude est le pentamère de la sous unité B de la toxine du choléra ($CtxB_5$).

L'interface de la $CtxB_5$ est découpée en plusieurs régions représentées graphiquement par le programme Gemini. A chaque région, Gemini associe un graphe qui décrit les interactions mises en jeu dans cette région d'interface. J'ai intéressé par la géométrie de l'interface β pour la mise en évidence de mécanismes d'influence en cascade.

L'objectif d'étude est de travailler sur cette interface β et de se concentrer en particulier sur les résidus impliqués dans le squelette (potentiel régulation de la structure secondaire) Y27, E29 et L31 et le résidu chargé au centre. La procédure utilisée dans cette analyse est le calcul des ΔG des mutations des résidus impliqués dans l'interface 23-31 avant et après mutation. Les résultats montrent que les doubles mutations L31F+Y27Q et L31F+E29D permettent de compenser le changement introduit par la mutation unique L31F. Ce qui suggère que les résidus Q27 et E29 bloque un phénomène engendré par la mutation L31F. De plus on peut aussi voir que le double mutant L31F+E29D est plus efficace pour empêcher la déstabilisation de L31F. Donc le phénomène de non additivité se produit par une communication pair à pair, créant un phénomène en cascade pour propager des changements structuraux nés de changements dans une séquence.

Les phénomènes d'influence ont été analysés par l'approche graphique. La mutation individuelle a un changement dans le nombre d'interaction plus que dans les doubles mutations. Ces résultats montrent l'existence de la communication pair à pair et que la régularisation de l'interface fait par des paires des mutations même si chaque mutations a un effet différent sur l'interface.

5.1 Cadre du problème

5.1.1 Le choix de l'interface β de la toxine du choléra

Dans des travaux présentés dans les Chapitres 7 et 8, j'ai étudié les caractéristiques de réseaux d'interfaces de géométrie β, c'est à dire des interfaces composées de deux brins β interagissant entre eux, à partir d'abord d'une base de données de 40 protéines oligomériques (chapitre 7) puis d'un set de 750 protéines oligomériques (chapitre 8). Cette géométrie est aussi observée dans de nombreuses protéines oligomériques impliquées dans des maladies neuro-dégénératives, la maladie de Alzheimer. Cependant les bases de données consistent en des cas non connus pour être impliquées dans ce type de maladie. Nous avons ainsi mis en évidence certaines caractéristiques communes à ces interfaces β, et pour cela j'ai choisi de considérer l'interface de géométrie β présente de $CtxB_5$ comme prototype d'étude pour la mise en évidence de mécanismes d'influence en cascade. L'interface de $CtxB_5$ est découpée en plusieurs régions (Figure 5.1) représentées graphiquement par le programme Gemini (chapitre méthodologie, Figure 3.2). A chaque région, Gemini associe un graphe qui décrit les interactions mises en jeu dans cette région d'interface en question, celui correspondant à l'interface β de $CtxB_5$ est indiqué en figure 5.2.

```
1EEI:D|PDBID|CHAIN|SEQUENCE
1TPQNITDLCA EYHNTQIYTL NDKIFSYTES LAGKREMAII TFKNGAIFQV
EVPGSQHIDS QKKAIERMKD TLRIAYLTEA KVEKLCVWNN KTPHAIAAIS MAN103
```

Figure 5.1 : La séquence primaire de $CtxB_5$ (PDB 1EEI). En rouge dans la séquence sont les différentes régions impliquées dans des interfaces (interactions intermoléculaires).

La région β de $CtxB_5$ implique des interactions entre les acides aminés 23 à 31 d'une chaîne et 96 à103 de la chaîne adjacente.

J'ai choisi de travailler sur cette interface β et de me concentrer en particulier sur les résidus impliqués dans le squelette (potentiel régulation de la structure secondaire) Y27, E29 et L31 et le résidu chargé au centre (E29). Comme mentionné précédemment, le but du travail est d'observer s'il existe des mécanismes de communications en cascade au sein de l'interface β entre ces résidus.

5.1.2 La communication pair à pair

Le monde de l'informatique est en effervescence autour d'un phénomène portant le nom de réseau pair à pair (Peer-to-Peer networks, P2P). Mal identifiée, mal comprise et mal considérée à ses débuts, l'idée de réseau pair à pair a beaucoup mûri aux cours des deux dernières années. Aujourd'hui, on parle de la communication pair à pair (communication de proche en proche) comme un modèle de communication capable de changer radicalement certaines approches de l'informatique en réseau. Ce concept introduit une relation d'égal à égal entre deux domaines. Dans son essence, l'informatique paire à paire se définit comme le partage des ressources et des services par échange direct entre systèmes. Ces échanges peuvent porter sur les informations, les cycles de traitement, la mémoire cache ou encore le stockage sur disque des fichiers. Dans le cadre de mon travail de thèse, on parle d'une communication paire à paire au sein d'une interface entre deux chaînes protéiques. Ici pair signifie acide aminé, et communication pair à pair signifie communication entre acides aminés de proche en proche afin de réguler des mécanismes structuraux tels que repliements et/ou assemblage).

Les caractéristiques du pair à pair

Un vrai système pair à pair se reconnaît par deux caractéristiques mises en évidence en posant les questions suivantes :

• Est-ce que la protéine permet à chaque pair (chaque acide aminé) de se connecter de manière intermittente avec d'autres acides aminés de façon variable, sans pour autant compromettre le bon fonctionnement du réseau? Ici, il s'agira par exemple de tester si le réseau résiste à des mutations de hots spots, mutations affectant la connectivité locale.

• Est-ce que la protéine donne à chaque pair une autonomie significative?

Si la réponse est oui à ces deux questions, le système est pair à pair. Si la réponse est négative aux deux questions, une autre manière de distinguer un système pair à pair est de raisonner en termes de "propriété de réseau". Il faut remplacer la question "Est-ce que le système donne à chaque pair une autonomie significative ?" par la question "Qui possède les ressources qui font tourner le système ?". Par exemple, quels sont les liens minimums à maintenir pour conserver les propriétés structurales d'une interface, c'est-à-dire le réseau minimum capable de résister à des mutations.

Une autre caractéristique du réseau pair à pair est que la qualité et la quantité des données disponibles augmentent à mesure que le nombre d'utilisations augmente. Dans le cadre des protéines, il pourrait s'agir d'un réseau capable de s'adapter à des mutations en évoluant en fonction des conséquences des mutations en termes de connections.

Ces différentes propriétés des réseaux pair à pair seront explorées tout au long de mon manuscrit de thèse afin d'argumenter sur la pertinence de décrire une protéine par un réseau pair à pair.

5.1.3 Réseaux d'interaction

Le réseau d'interactions généré par Gemini décrit l'interface β de $CtxB_5$ en deux sous réseaux (Figure 5.2). Un réseau constitué des interactions entre atomes du squelette des acides aminés et un réseau constitué des interactions entre atomes de leur chaîne latérale.

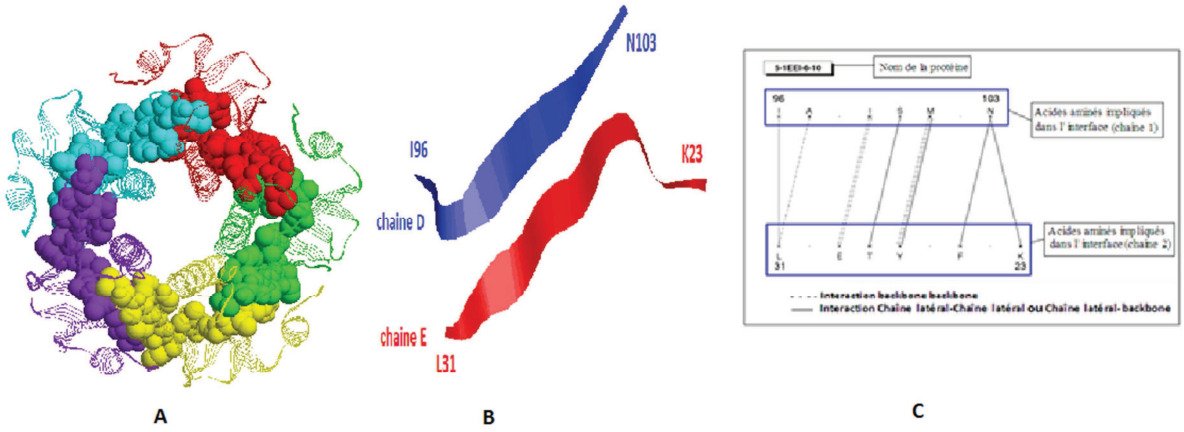


Figure 5.2 : A. Exemple d'interface dans $CtxB_5$. Structure atomique du pentamère de la toxine du choléra. Chaque chaîne est indiquée par une couleur différente. La région d'interface composée des résidus 23 à 31 et 96 à 103 est illustrée par des sphères. B. **La région d'interface composée des résidus 23 à 31 et 96 à 103. C. Réseau d'interactions généré par Gemini pour la région d'interface β composée des résidus 23 à 31 et 96 à 103.** La ligne du bas correspond aux résidus 23 à 31 d'une chaîne et la ligne du haut correspond aux résidus 96 à 31 de la chaîne adjacente. Le réseau d'interaction est composé de deux sous réseaux atomiques distingués par des traits continus (interactions entre atomes des chaînes latérales) et des traits discontinus (interactions entre atomes du squelette).

5.2 Protocoles

Pour tester la possibilité de communication en cascade, j'ai généré les mutations *in silico* des résidus Y27, E29 et L31 du brin β composé des résidus 23 à 31 de l'interface et j'ai comparé les réseaux avant et après mutations (Gemini et Spectral-Pro) ainsi que les énergies d'interaction.

J'ai suivi ce protocole :

La première étape était de générer le graphe Gemini de la région 96-103 ; 23-31, ce graphe permet de caractériser l'interface β comme un sous-réseau des hots spots représentant l'interface β entre les deux chaînes adjacentes D et E. La deuxième étape était de calculer les énergies d'interaction en utilisant l'algorithme Fold-X (voir chapitre 3, méthodologie). La troisième étape était d'utiliser Spectral-Pro pour comparer les propriétés de connectivités des différents réseaux.

5.3 Résultats

Je présente dans cette section les résultats obtenus sur le réseau d'interaction étudié de la région β (96-103 ; 23-31). Ces résultats sont présentés en utilisant des fichiers PDB muté individuellement, doublement et triplement des trois acides aminés Y (tyrosine), E (acide glutamique) et L (leucine) dans les positions 27, 29 et 31 respectivement.

Brin β 23 à 31	**Energie d'interaction (Kcal/mol)**	**$E_{mutation}$-$E_{origine}$ (Kcal/mol)**	**Additivé**	**Réseau global**		**Réseau local**	
				Poids	**Degré**	**$Poids_{muté}$-$poids_{origine}$**	**$Degré_{muté}$-$Degré_{origine}$**
KIFSYTESL	-11	na		1474	182	na	na
KIFSQTESL	-9	2		1454	180	0	-1
KIFSYTDSL	-12	-1		1394	182	-37	-1
KIFSYTESF	0	11		1536	180	35	0
KIFSQTDSL	-11	0	Oui	1386	178	Q27 : -6 D29 : -37	Q27 : -2 D29 : -1
KIFSYTDSF	-4	7	Non	1480	180	D29 : -35 F31 : 28	D29 : -1 F31 : 0
KIFSQTESF	-7	4	Non	1464	182	Q27 : -5 F31 : 23	Q27 : -2 F31 : 0
KIFSQTDSF	-3	8	Non	1366	170	Q27 : 0 D29 : -38 F31 : 26	Q27 : -1 D29 : -1 F31 : 0

Tableau 5.1: les énergies d'interaction après mutation des résidus étudiés (en rouge). La première ligne indique la séquence d'origine sans mutation.

Les résultats des calculs énergies d'interaction sont indiqués dans le tableau 5.1, colonne 2, la différence d'énergie avant et après mutation, colonne 3, le poids global du réseau (somme de toutes les interactions atomiques), colonne 5 et le degré global, colonne 6 (somme des paires d'acides aminés en interaction). La colonne 4 porte l'information d'additivité des mutations, c'est à dire si l'effet des mutations est diffèrent lorsque les mutations sont combinées (non additif) ou non (voir chapitre 4). Les différences entre les

paramètres poids et degré globaux muté et natif indiquent des changements structuraux résultant des mutations (Perte ou gain d'interactions ou de pairs d'acides aminés).

Dans la première colonne du tableau est indiquée la séquence du brin β composé d'une des résidus 23 à 31. Les résidus Y27 E29 et L31 sont mutés par Q (glutamine), D (acide aspartique) et F (phénylalanine), respectivement. Ces mutations conservent la propriété chimique des acides aminés d'origine, mais pas nécessairement la propriété géométrique (Y27Q et L31F). La mutation des acides aminés Y27 et E29 affecte peu l'énergie d'interaction au contraire de la mutation de L31. De la même manière la mutation double Y27Q et E29D n'affecte pas l'énergie d'interaction alors que toutes les mutations combinées avec celle du résidu L31 l'affectent. Toutes ces mutations multiples ont des effets non additifs sur l'énergie d'interactions (Tableau 5.1).

Plus précisément, on observe des phénomènes compensatoires puisque les mutations Y27, E29 et F31 déstabilisent moins que la mutation individuelle F31. Cela suggère que les acides aminés du réseau squelette communiquent entre eux puisque les effets de leur mutation ne sont pas indépendants. En d'autres termes l'interface β de $CtxB_5$ n'est pas seulement décrite par un réseau d'interactions mais fonctionne comme tel.

Le but de notre étude était de voir s'il existait des chemins de communication entre des acides aminés éloignés (communication à longue distance) au sein de l'interface. Les résultats des ΔG du tableau montrent que les doubles mutations L31F+Y27Q et L31F+E29D permettent de compenser le changement introduit par la mutation unique L31M. Ce qui suggère que les résidus Q27 et E29 bloquent un phénomène engendré par la mutation L31F. De plus on peut aussi voir que le double mutant L31F+E29D est plus efficace pour empêcher la déstabilisation de L31M.

La figure 5.3 montre que la mutation individuelle L31F perd l'interaction avec le résidu V50 et ajoute une nouvelle interaction avec A95. Par contre, quand le résidu L31 est muté en combinaison avec le résidu Y27 ou avec le résidu E29, on remarque l'apparition d'un nouveau lien entre les résidus 31 et 57 et la disparition de lien avec le résidu A95. Le lien avec le résidu V50 n'est pas récupéré dans les doubles mutations mais remplacé par le lien avec le résidu 57 dans le voisinage. On peut expliquer cet effet d'influence comme un

changement structural allant de résidus en résidus (ex. du résidu 27 au résidu 29 et au résidu 31).

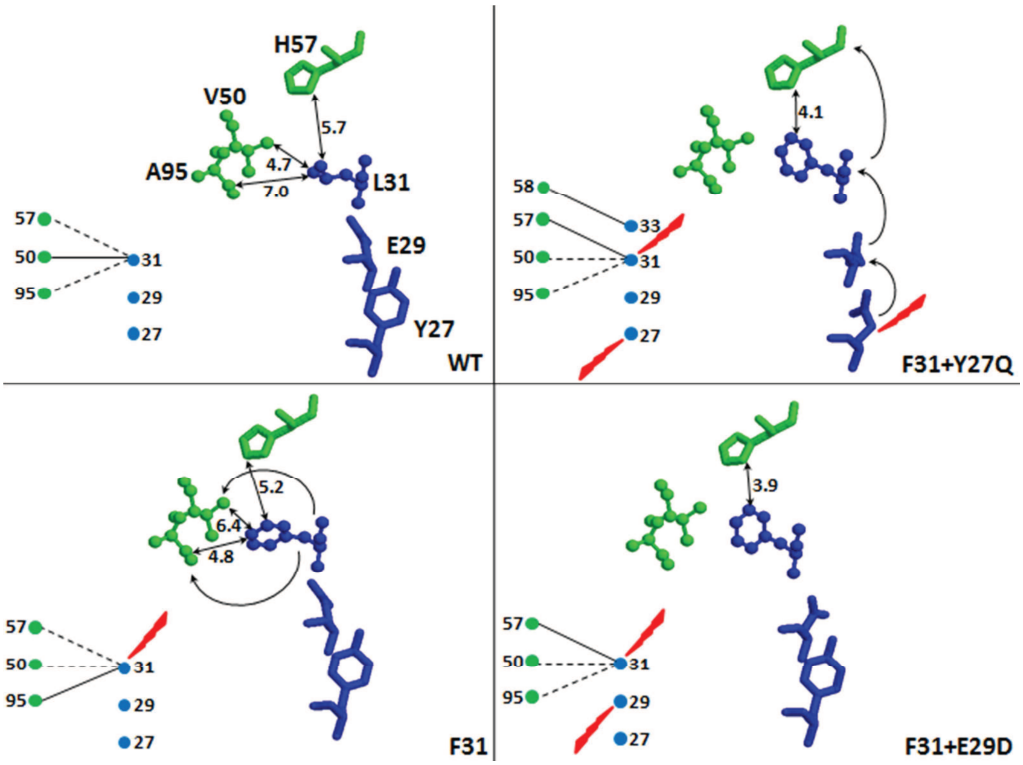


Figure 5.3 : représentation des acides aminés étudiés en mutation individuel F31 et double F31+Q27 et F31+D29. Les traits continus indiquent l'existence d'une interaction entre les résidus et l'absence de l'interaction est montrée par un trait discontinu.

On peut voir en figure 5.4, que les effets d'influence (non additifs) ne sont pas systématiques puisque les mutations individuelles Y27Q et E29D perturbent les mêmes interactions individuellement qu'ensemble (contrôle négatif). On a bien un effet additif dans cette situation. Il est important de remarquer que ces mutations entraînent des pertes d'interactions sans pour autant affecter de façon significative l'énergie d'interaction. Le résidu 31 a aussi des interactions avec les résidus 64, 65 et 68 qui sont proches des résidus 64, 67 et 71 en interaction avec les résidus 27 et 29. Il est donc possible que les réseaux avant et après mutations soit différents mais équivalents, n'entrainant pas de modification de l'énergie d'interaction. Cette possibilité serait en bon accord avec la propriété des réseaux pair à pair qui permettent d'avoir plusieurs combinaisons de liens pour chaque pair sans pour autant perturber l'intégrité du réseau. En d'autres termes il existerait plusieurs réseaux alternatifs, c'est-à-dire une certaine plasticité de modifications structurales neutres en termes de propriétés de réseau.

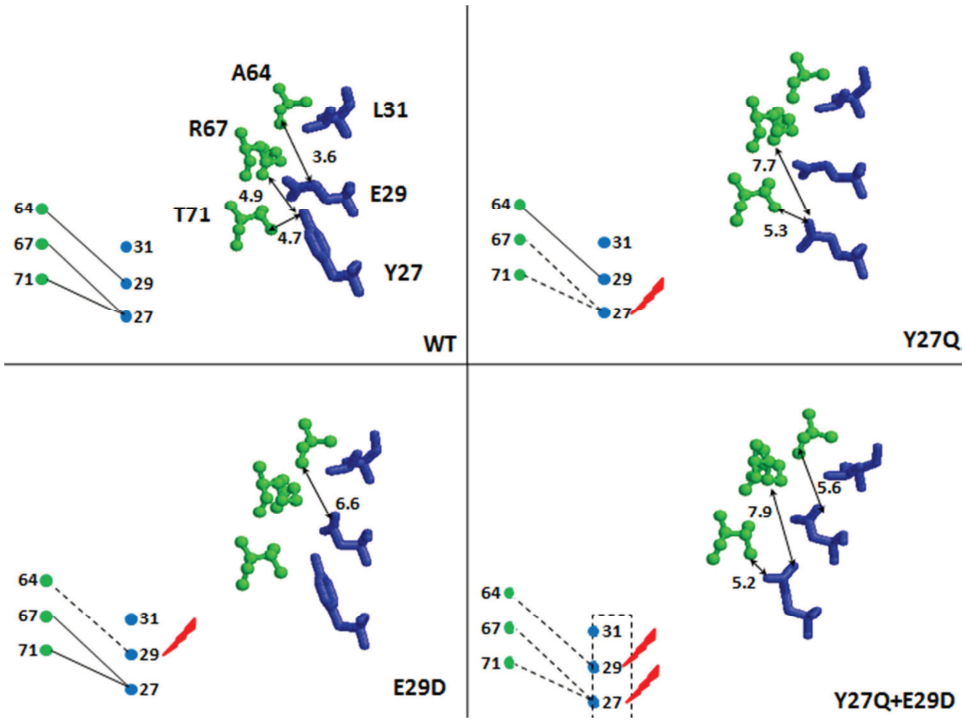


Figure 5.4 : représentation des acides aminés étudiés en mutation individuel Q27 et D29 et double Q27+D29.

Indépendamment des mutations combinées, la communication pair à pair peut aussi être mise en évidence en comparant les différences de degré et poids globaux avec les différences de degrés et poids locaux. Le calcul du rapport entre degré et poids globaux et locaux est supérieur à 2, ce qui signifie qu'on a un effet au-delà du local. La communication entre pair renforce le fait que la régulation structurale se fait à travers des pairs de résidus plutôt qu'au niveau d'un résidu individuel.

Peut-on trouver une explication des phénomènes d'influence en utilisant l'approche des graphes ? Pour répondre à cette question, j'ai analysé les résultats des mutations étudiées tels que les mutations neutres Y27Q, E29D et Y27Q+E29D (Figure 5.5) qui n'ont que des effets de changements de poids, puis les mutations influencent avec des effets sur les poids et sur les paires d'acides aminés (Figure 5.6). Pour cela, j'ai additionné tous les poids des paires d'une même région d'interface.

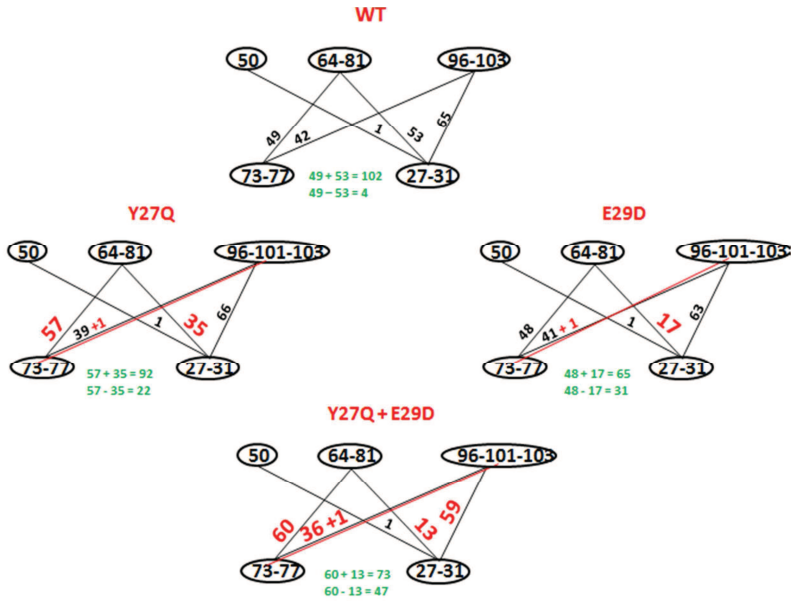


Figure 5.5 : les réseaux d'interactions dans les mutations individuelles et doubles.

Les mutations neutres: Y27Q, E29D et Y27Q+E29D montrent qu'ils ont même réseaux que la native, la région 27-31se détache de la région de 64- 81 tandis que la région 74-77 attache plus à lui. L'ajout et perd des liens de la même région est un effet remarquable après la mutation L31F. Pour les mutations QF, DF, QDF, l'effet des mutations est lié à la mutation de L31F en enlevant le lien avec le résidu A95, encore l'élimination du lien avec le résidu V50 et de plus de résidu H57 mais qui a même région d'interaction dans le réseau de liaison (58, 33). Alors que la mutation QF suggère un ajout de deux nouveaux liens que la native (74, 74) et (29, 101). La double mutation commence à se stabiliser la protéine en regardant le réseau d'interaction, l'absence du lien avec le résidu V50 et l'ajout du résidu H57 sont les seules changements apparait. La double mutation DF montre qu'aucun lien de backup, alors peut-être ΔG correspond à l'énergie après réparation partielle de la mutation L31F c'est-à-dire qu'on voit beaucoup d'interactions déstabilisent la protéine : dans les mutations QF, DF et QDF absence du V50 et ajout du H57 dans les trois mutations mais plus quand diminue le nombre de liens l'énergie de l'interaction augmente donc stabilise. En fait trop de lien, trop de connectivite implique augmentation des effets d'influences et de la propagation : danger (voir le chapitre 8 cas de la p53). Dans les régions 27-31 et 74-77 détachés 64-81, la triple mutation QDF montre une forte diminution des interactions.

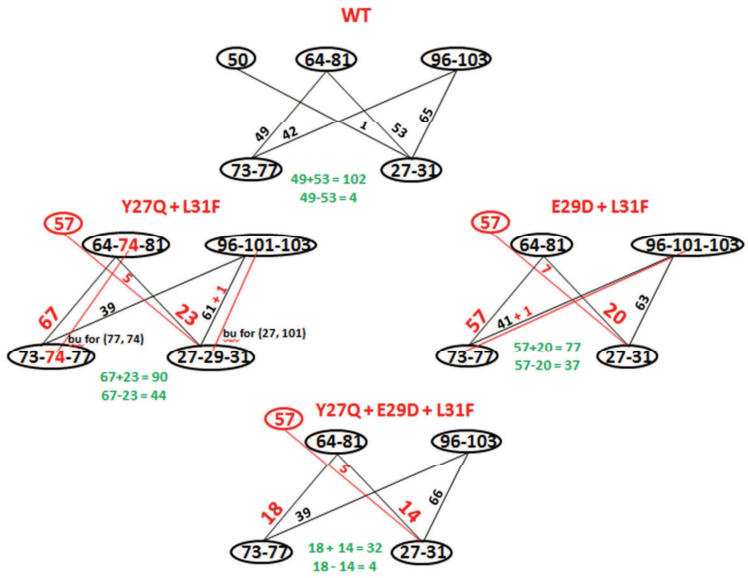


Figure 5.6 : les réseaux d'interactions dans les mutations individuel, double et triple. Communauté 50 en moins, communauté 57 introduite.

5.4 Conclusion

La mesure globale de graphe mesure bien la communication dans un réseau intra-inter moléculaire puis le réseau inter-inter moléculaire. Le phénomène de non additivité s'opère par une communication pair à pair, produisant un phénomène en cascade pour propager des changements structuraux nés de changements dans une séquence. Les résultats des mutations des résidus L31F+Y27Q et L31F+E29D montrent l'existence de la communication pair à pair tel que chaque mutation individuelle a un changement dans le nombre d'interactions plus que dans les doubles mutations, donc l'interface est régulée par des paires des mutations même si chaque mutation a un effet différent sur l'interface. Cependant, notre étude permet aussi de montrer des mutations neutres ou le réseau d'interactions après les mutations est similaire à celui de la protéine native, c'est le contrôle négatif qui est présenté dans le cas des mutations individuelles et doubles de Y27Q et E29D qui montre l'additivité des mutations individuelles et la double mutation.

Chapitre 6: Peut-on anticiper le mécanisme d'assemblage d'une protéine par des approches réseaux ?

Les résultats précédents ont montré que la régulation de la formation d'une interface se fait par une communication entre les résidus en dehors de l'interface et les résidus de l'interface par un processus en cascade. L'approche réseau m'a permis de mettre en évidence des communications entre les réseaux intramoléculaires et intermoléculaires. Ces résultats nous encouragent à anticiper à savoir le mécanisme d'assemblage des protéines. Les chemins de communication ont été testé dans la $CtxB_5$ et ont suggéré que cette communication est peut être responsable de la coordination entre les étapes de repliement (intra) et les étapes d'association du mécanisme d'assemblage. Donc, j'ai étudié quels étaient les résidus impliqués dans la communication intra/inter pour deux toxines (la toxine du choléra ($CtxB_5$) et la toxine labile (LTB_5)), sachant que ces deux toxines suivent des mécanismes d'assemblage différents et qu'elles ont une identité de 87% de séquence et elles suivent des mécanismes d'assemblage différents.

Les résidus impliqués dans la communication intra/inter pour les deux toxines ont été analysés après le calcul des énergies d'interaction et de la stabilité pour chaque mutation individuelle des acides aminés de la chaîne des deux toxines. Si la communication inter/intra est responsable de la coordination des étapes de repliement et d'association, et que les deux toxines ne suivent pas le même mécanisme d'assemblage une se replie entièrement et s'associe alors que l'autre se replie et s'associe simultanément, on peut s'attendre à une distinction dans la coordination et donc dans les résidus impliqués dans la communication intra/inter.

Comme indication des résidus impliqués dans la communication entre les réseaux intramoléculaires et intermoléculaires, j'ai compte le nombre de résidus dont la mutation affectait à la fois l'énergie d'interaction et la stabilité de la toxine. D'après les résultats, la mutation de 29 résidus sur 103 affecte à la fois l'énergie d'interaction et la stabilité du pentamère de la toxine du choléra alors que la mutation de seulement 19 résidus affecte à la fois l'énergie d'interaction et la stabilité du pentamère de la thermolabile entereotoxine LTB5. Un tiers des résidus (10 résidus) concernés dans $CtxB_5$ sont impliqués dans des interactions intermoléculaires entre régions d'interface différentes pour seulement un sixième des résidus (3 résidus) concernés dans LTB_5.

Ces résultats ont montré que la formation des interfaces se fait par deux réseaux différents : réseau d'interactions entre différentes régions ou les interactions

intermoléculaires connectent des résidus de régions d'interface différentes (cas de la $CtxB_5$) et un réseau d'interactions à l'intérieur d'une seule région ce qu'il suggère moins de connexion entre les différentes régions (cas de LTB_5).

6.1 Cadre de problème

6.1.1 Mécanisme d'assemblage protéique

Le repliement des protéines peut être analysé de deux façons distinctes : l'une informative, contenue dans le code de la séquence et l'autre mécanique, avec la description de la cinétique et de la thermodynamique des processus qui décrivent l'évolution conformationelle de la protéine au cours du repliement. Pour comprendre le chemin de repliement des protéines on pourrait prédire la structure 3D d'une protéine à partir de sa séquence primaire, créer des protéines plus stables qui se replient plus facilement, identifier les résidus essentiels pour le repliement et comprendre les maladies liées à l'altération de la conformation des protéines.

Pour aborder le problème, j'ai choisi de comparer les propriétés de réseau des deux toxines AB_5, $CtxB_5$ et LTB_5. Les pentamères de la toxine thermolabile B_5 (LTB_5) et de la toxine cholérique B ($CtxB_5$) partagent une identité de séquence de 82 % (94 % d'homologie de séquence) et des structures atomiques presque superposables mais néanmoins s'assemblent par deux mécanismes différents. D'après les résultats des chapitres 4 et 5, cette différence de mécanisme d'assemblage pourrait être due à des différences de composition en acide aminés menant à des différences de voisinages créant des chemins de communication différents.

6.1.2

6.1.3 Type de mécanisme d'assemblage

La toxine du choléra suit un mécanisme d'assemblage de type « fly-casting » ou les réactions de repliement et d'assemblage sont concomitants [2]. Le repliement partiel du monomère $CtxB_5$ est pH dépendent, à pH acide la toxine n'acquiert pas le repliement partiel nécessaire à son bon assemblage. Au contraire LTB_5 se replie très rapidement et son assemblage a une pH-dépendance autour de 7.0. La deuxième étape est la fixation du brin β dans la position correcte pour s'assembler, cette étape est sous le contrôle de la trans-cis isomérisation de la proline 93 [22]. Les résultats des analyses réseaux (chapitre 5) permettent d'émettre l'hypothèse que les résidus 94, 93, 92 jouent un rôle dans la formation d'une

interface avec les résidus de 1 à 3 d'une autre chaîne (Figure 6.1). La formation de cette interface pourrait rigidifier le mouvement du brin β composé des résidus 96 à 103 et permettre dans une troisième étape la formation de l'interface avec le brin β composé des résidus 23 à 31 d'une troisième chaîne (Figure 6.1). Contrairement à $CtxB_5$, les monomères de LTB_5 se replient quasiment sous une forme native avant de s'associer entre eux [23].

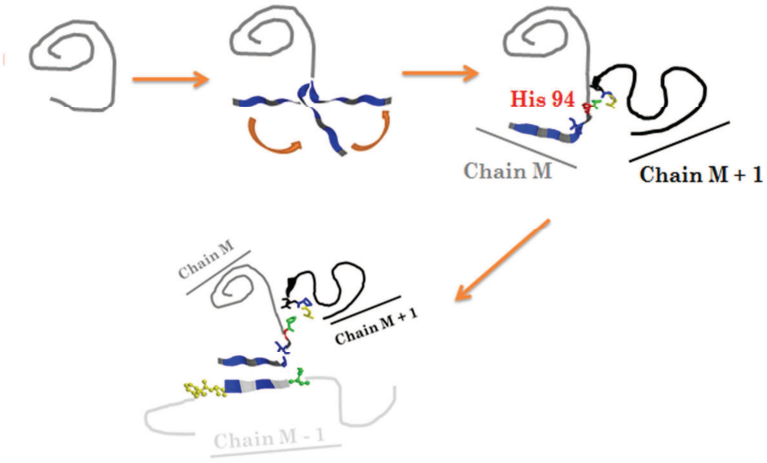


Figure 6.1 : les étapes suivies pour le mécanisme d'assemblage de la toxine du choléra

6.1.4 Types d'interaction

Les interactions protéines sont un aspect essentiel des processus biologiques. Elles sont fortement impliquées dans la formation de structures macromoléculaires, dans la signalisation, dans la régulation et dans les différentes voies métaboliques.

Les **interactions intra et intermoléculaires** sont toutes des interactions non covalentes (interactions électrostatiques, liaisons hydrogènes, forces de Van Der Waals, effet hydrophobe). C'est la somme de ces multiples interactions de faible énergie qui confère leur stabilité, mais aussi leur flexibilité et leur dynamique puisque ces interactions peuvent fluctuer rapidement sous l'effet de la température, du pH et de la force ionique.

6.2 Protocole

Dans ce chapitre, j'ai comparé des deux toxines $CtxB_5$ et LTB, et comment on peut anticiper le mécanisme d'assemblage d'une protéine en utilisant des approches réseaux. Pour cela, j'ai muté tous les acides aminés existant dans les chaînes protéiques un par un pour les deux toxines par l'acide aminé asparagine (acide aminé a des propriétés chimiques moyennes). Cette étude est faite par l'analyse des résultats des énergies d'interaction, stabilité

et la géométrie de chaque mutation étudiée en utilisant l'algorithme Fold-X et le programme Spectral-Pro.

Pour chaque toxine :

- Muter tous des acides aminés un par un par l'acide aminé asparagine.
- Calculer l'énergie d'interaction et la stabilité (définir les énergies) pour chaque mutation de chaque acide aminé.
- Calculer le nombre des interactions globales atome-atome (poids) et résidu-résidu (degré) pour chaque mutation de chaque acide aminé.

Ces étapes nous permettons d'explorer des pistes pour comprendre ce qui détermine le mécanisme d'assemblage.

6.3 Résultats

Les résultats du chapitre 4 ont montré qu'il existe des chemins de communication entre les résidus en dehors de l'interface et les résidus de l'interface, et que cette communication est peut être responsable du mécanisme d'assemblage. Pour tester cette hypothèse plus avant, j'ai étudié quels étaient les résidus impliqués dans la communication intra/inter pour les deux toxines (la toxine du choléra ($CtxB_5$) et la toxine labile (LTB_5)), puisque ces deux toxines suivent des mécanismes d'assemblage différents.

Cette étude se base sur les calculs des énergies d'interaction (intermoléculaire) et de la stabilité (intramoléculaire) des deux toxines, pour les toxines de type sauvage ainsi que pour toutes les mutations individuelles de tous leurs acides aminés par l'acide aminé asparagine. L'analyse des résultats se fait par la comparaison du nombre de mutations qui affectent à la fois l'énergie d'interaction et la stabilité de la toxine, comme indication d'un résidu impliqué dans la communication entre les réseaux intramoléculaires et intermoléculaires. Les résultats sont indiqués dans les tableaux 6.1 et 6.2, respectivement.

$CtxB_5$ (1EEI)		
Mutants	**Energie d'interaction (Kcal/mol)**	**Energie de stabilité (Kcal/mol)**
WT	-11	8
K69N	-17	13
K63N	-15	10
D70N	-15	10
E36N	-15	12
E79N	-14	7
K43N	-14	10
Y12N	-14	12
E83N	-14	7
D7N	-14	12
N103K	-14	6
K23N	-13	11
K34N	-13	10
R73N	-13	11
F25N	-13	10
A80N	-13	10
V50N	-12	15
L8N	-12	12
Y76N	-12	14
S26N	-10	10
L77N	-10	11
M37N	-10	14
A32N	-10	11
M101N	-10	15
Q61N	-9	11
Y27N	-9	18
S30N	-9	11
L31N	-9	11
P93N	-8	24
A98N	-5	14

Tableau 6.1 : les mutations affectent à la fois l'interface et la stabilité de la toxine du choléra ($CtxB_5$)

Le tableau présente les résultats des mutations qui affectent l'interface et la stabilité de la toxine du choléra, tel que la première colonne est le nom des mutants (acide aminé avant la mutation, la position, l'acide aminé après mutation), la deuxième colonne est le calcul des énergies d'interactions pour chaque mutation et la troisième présente le calcul des énergies de la stabilité globale de la protéine pour chaque mutation.

Les mêmes étapes sont faites pour la deuxième toxine (LTB), les résultats sont illustrés dans le tableau 6.2 suivant :

1LTB		
Mutants	**Energie d'interaction (Kcal/mol)**	**Energie de stabilité (Kcal/mol)**
WT	-6	25
T78N	-13	27
E79N	-12	27
S55N	-11	27
I58N	-11	27
D70N	-10	30
Y76N	-10	30
I99N	-10	32
Q61N	-10	28
R35N	-10	37
K62N	-10	34
K63N	-9	27
V50N	-9	28
H57N	-8	33
W88N	-8	32
K69N	-8	27
E11N	-8	29
D22N	-8	27
D59N	-8	30
M101N	-4	30

Tableau 6.2 : les mutations affectent à la fois l'interface et la stabilité de la toxine labile (LTB)

D'après les résultats, la mutation de 29 résidus sur 103 affecte à la fois l'énergie d'interaction et la stabilité du pentamère de la toxine du choléra $CtxB_5$ alors que la mutation de seulement 19 résidus affecte à la fois l'énergie d'interaction et la stabilité du pentamère de la thermolabile entereotoxine LTB_5. Un tiers des résidus concernés dans $CtxB_5$ forment des interactions intermoléculaires entre régions d'interface différentes (c'est-à-dire 10 résidus) pour seulement un sixième des résidus concernés dans LTB_5 (3 résidus).

En figure 6.2, j'ai résumé ce dernier résultat sur un schéma simplifié des interactions des deux toxines. Les X sont des régions d'interface différentes et les liens sont les interactions entre deux chaînes protéiques adjacentes. Dans le cas de $CtxB_5$, on a des interactions partagées entre différentes régions (Figure 6.2a) et dans le cas de LTB_5, les

interactions se font à l'intérieur d'une seule région (Figure 6.2b). Dans un réseau de type 6.2a, les interactions intermoléculaires connectent des résidus de régions d'interface différentes. Une telle connectivite va nécessairement entraîner le rapprochement dans l'espace de résidus éloignés dans la chaîne (Figure 6.3) et donc influer sur le repliement de la chaîne. Dans le réseau de type 6.2b, la connectivité se fait entre résidus proches dans la chaîne et on ne s'attend donc pas à un effet des interactions intermoléculaires sur le repliement intramoléculaire de la chaîne. En fait ce type de réseau correspondrait plutôt bien à une situation ou le repliement et donc la position spatiale des résidus est déjà conçue avant l'étape d'association, de telle sorte qu'on observe des interactions locales fortes plutôt que des interactions à longue distance.

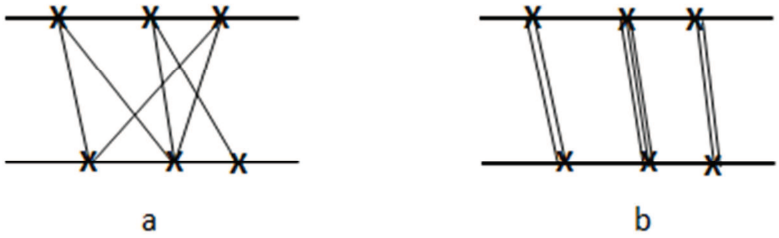


Figure 6.2 : représentation des interactions entre deux chaînes protéiques. a : plus de pair (hot spot) dans des triangles à l'intérieur d'un domaine (cas de la toxine du choléra). b : moins de pair (hot spot) dans des triangles à l'intérieur d'un domaine (cas de la toxine labile).

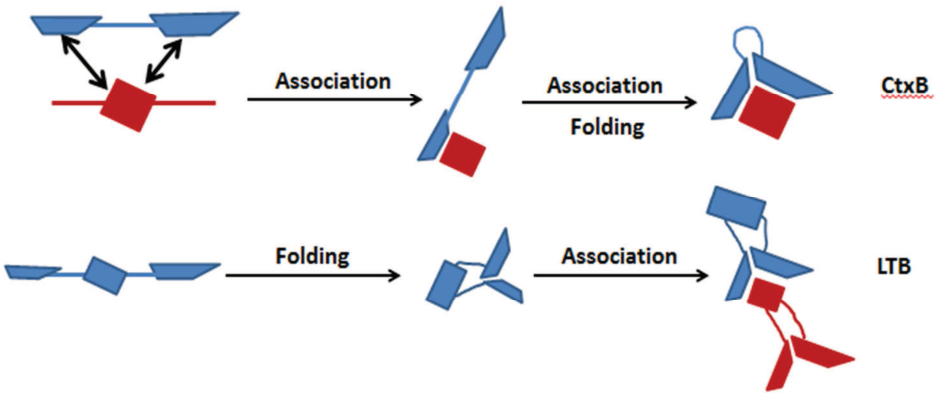


Figure 6.3 : Différentes étapes du mécanisme d'assemblage pour les toxines CtxB et LTB

Il semble que lors d'un assemblage ou les étapes d'association et de repliement sont concomitantes, on peut s'attendre à avoir plus de résidus impliqués à la fois dans les interactions intermoléculaires et intramoléculaires ($CtxB_5$) et aussi dans la coordination entre régions d'interface ($CtxB_5$). Par contre lorsque le repliement est fait avant l'association, les résidus impliqués dans les interfaces ont déjà adoptés les positions spatiales adéquates pour

l'association, on peut s'attendre à moins de coordination entre les étapes d'association et de repliement et a une plus grande distinction entre les résidus des réseaux intramoléculaires et intermoléculaires respectifs, comme observes pour la LTB_5.

Les résultats permettent de trouver une cohérence entre les mécanismes d'assemblage mis en évidence expérimentalement et des propriétés de réseaux des deux toxines. Ceci n'était pas évident au vue des structures atomiques très proches des deux toxines. Cette approche permet d'ouvrir des pistes pour appréhender les mécanismes d'assemblage à partir des structures atomiques des protéines.

Chapitre 7: Article publie: Beta-Strand Interfaces of Non-Dimeric Protein Oligomers Are Characterized by Scattered Charged Residue Patterns

Une base de données constituée de 40 protéines oligomériques ayant toutes au moins une interface β, interface faite de l'association de deux brins β a été étudiée dans ce chapitre. Les oligomères étudiés sont tous de stœchiométrie supérieure à deux et inférieur à 13. Les interfaces β sont modélisées par deux réseaux d'interactions : les interactions impliquant des atomes du squelette des acides aminés (réseau squelette) et les interactions impliquant des atomes de la chaine latérale (réseau chaine latérale).

Les réseaux de la chaîne latérale ont la spécificité d'avoir des résidus chargés aux extrémités et au centre de l'interface. Cette répartition de charge permet de les différencier des séquences de feuillets β intermoléculaires des dimères, des feuillets β intramoléculaires et des feuillets β formant des fibres de type amyloïde. Au contraire les réseaux squelettes partagent les mêmes acides aminés que les autres feuillets susmentionnés, la nature des acides aminés de ses réseaux correspond bien à celle attendue pour une structure secondaire de feuillet β.

A partir de l'analyse de la base de données, des tests expérimentaux ont été fait en utilisant le pentamère de la toxine du choléra ($CtxB_5$) comme prototype d'interface β. L'idée était d'étudier l'assemblage de $CtxB_5$ en pentamère *in vitro* puis d'ajouter dans la solution de réassemblage des peptides dont les séquences correspondaient à celles de l'interface β de $CtxB_5$. Si ces peptides sont capables de former une interface avec des monomères de $CtxB_5$, l'assemblage de la toxine en pentamère sera inhibé. Une fois la mise en place du protocole, les séquences des peptides ont été modulées en accord avec les résultats théoriques afin de tester si après mutation d'un acide aminé les peptides restaient ou non capable d'inhiber l'assemblage de la toxine en pentamère. La comparaison des résultats expérimentaux et théoriques permettront de mieux comprendre les éléments essentiels à la formation d'une interface β. Les résultats théoriques montrent des acides aminés communs à la géométrie β et des acides aminés permettant de fournir une certaine spécificité à l'interface, intermoléculaires de type dimère, intermoléculaires de type plus que dimère, de feuillets intramoléculaires ou encore de fibre. Le but de la partie expérimentale est de produire des peptides capables d'inhiber un assemblage sans avoir la séquence exacte d'une protéine mais plutôt une séquence générique de la géométrie de l'interface et une spécificité associée à un type de protéines telle que celle d'une fibre. Ce type de peptides pourrait permettre de cibler avec un seul compose des pathologies associées à un groupe de protéines tout en limitant les

risques de réactions aspécifiques. Ces peptides construits à partir de $CtxB_5$ pourront alors être testés dans leur capacité à inhiber la formation d'une interface β sur d'autres protéines. La première expérience était donc de vérifier si les peptides de type sauvage pouvaient former une interface avec $CtxB_5$, puis j'ai réussi à tester un peptide muté. Les résultats obtenus ont montré que lorsque les brins des interfaces β sont produits individuellement comme des peptides synthétiques, ils sont capables d'inhiber l'assemblage de la toxine même dans une version mutée.

Les deux réseaux d'interactions formant une interface β ont donc chacun leurs propres caractéristiques qui peuvent être associés à un rôle distinct dans la formation de l'interface.

OPEN ACCESS Freely available online

PLoS one

Beta-Strand Interfaces of Non-Dimeric Protein Oligomers Are Characterized by Scattered Charged Residue Patterns

Giovanni Feverati[1], Mounia Achoch[1,2], Jihad Zrimi[1,2], Laurent Vuillon[3], Claire Lesieur[1,4]*

1 Université de Savoie, Annecy le Vieux Cedex, France, **2** Laboratoire de Chimie Bioorganique et Macromoléculaire (LCBM), Faculté des Sciences et Techniques-Guéliz, Université Cadi Ayyad, Marrakech, Maroc, **3** LAMA, Université de Savoie, Le Bourget du Lac, France, **4** AGIM, Université Joseph Fourier, Archamps, France

Abstract

Protein oligomers are formed either permanently, transiently or even by default. The protein chains are associated through intermolecular interactions constituting the protein interface. The protein interfaces of 40 soluble protein oligomers of stœchiometries above two are investigated using a quantitative and qualitative methodology, which analyzes the x-ray structures of the protein oligomers and considers their interfaces as interaction networks. The protein oligomers of the dataset share the same geometry of interface, made by the association of two individual β-strands (β-interfaces), but are otherwise unrelated. The results show that the β-interfaces are made of two interdigitated interaction networks. One of them involves interactions between main chain atoms (backbone network) while the other involves interactions between side chain and backbone atoms or between only side chain atoms (side chain network). Each one has its own characteristics which can be associated to a distinct role. The secondary structure of the β-interfaces is implemented through the backbone networks which are enriched with the hydrophobic amino acids favored in intramolecular β-sheets (MCWIV). The intermolecular specificity is provided by the side chain networks via positioning different types of charged residues at the extremities (arginine) and in the middle (glutamic acid and histidine) of the interface. Such charge distribution helps discriminating between sequences of intermolecular β-strands, of intramolecular β-strands and of β-strands forming β-amyloid fibers. This might open new venues for drug designs and predictive tool developments. Moreover, the β-strands of the cholera toxin B subunit interface, when produced individually as synthetic peptides, are capable of inhibiting the assembly of the toxin into pentamers. Thus, their sequences contain the features necessary for a β-interface formation. Such β-strands could be considered as 'assemblons', independent associating units, by homology to the foldons (independent folding unit). Such property would be extremely valuable in term of assembly inhibitory drug development.

Citation: Feverati G, Achoch M, Zrimi J, Vuillon L, Lesieur C (2012) Beta-Strand Interfaces of Non-Dimeric Protein Oligomers Are Characterized by Scattered Charged Residue Patterns. PLoS ONE 7(4): e32558. doi:10.1371/journal.pone.0032558

Editor: F. Gisou van der Goot, Ecole Polytechnique Federale de Lausanne, Switzerland

Received October 17, 2011; **Accepted** January 29, 2012; **Published** April 9, 2012

Funding: This work was supported by The system complex Rhone Alpes IXXI (5000 euros). Supported by the University of Savoie (4000 euros)The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: lesieur@lapp.in2p3.fr

Introduction

Most proteins are made of more than one polypeptide chain to carry out their biological function [1,2]. They are referred to as protein oligomers and have what is called a quaternary structure. In addition, numerous monomeric proteins associate transiently in binary or in higher stœchiometries (number of chains associated in a protein oligomer) during their life span. The formation of protein oligomer, known as protein assembly, is also a common reaction used by pathogens to produce killing "machineries". One good example is the pore forming toxins produced by pathogenic bacteria such as *Bacillus anthracis*, *Staphylococcus aurus* and *Aeromonas hydrohilae*. This mechanism is also responsible for protein misfolding diseases through the production of "amyloid" oligomers and fibers (e.g. Alzheimer, Parkinson, Creuzfeld Jacob) [3,4,5,6,7,8,9].

Intermolecular contacts (contacts between chains) exist only in multiple chain proteins. These contacts constitute what is called the protein interface and are formed through particular interaction patterns. Unfortunately, despite extensive analyses, the identification of the patterns responsible for permanent contacts remains difficult. This is due to the broad diversity of the contact solutions [10,11]. The rationalization of known patterns of protein interfaces is also far from accomplished.

The patterns result from geometrical and chemical complementarities between the two partners. Numerous reports on protein interfaces, based on theoretical and experimental approaches, allow understanding some of the general rules underlying intermolecular contacts (for reviews see [2,10,12]).

First, one needs to distinguish within the interface, the amino acids involved in intermolecular contacts, the so called "hot spots", from those who are not. Several programs can identify theoretical hot spot residues at interfaces based on: (i) distance cuts-off combined or not with some chemical selection, (ii) solvent accessible surfaces, (iii) geometrical selection (e.g. Voronoi cells) or (iv) evolutionary conserved residues [2,13,14,15]. All require the atomic structure of the protein oligomer. Experimental evidences have also confirmed the presence of hot spot residues in interfaces (for review see [2]). One beautiful example is the selective effect of the mutation of only some of the residues of the interface on the protein assembly of the heptameric co-chaperone cpn10 [16].

Second, the interaction patterns of protein interfaces are related to their secondary and tertiary structures as it was initially described by Sir Francis Crick for α-coiled interfaces with the discovery of the heptaed sequences [17,18,19,20,21,22,23,24]. The importance of the structure of the interface in the implementation of a particular motif has been now generalized with high-throughput interaction discovery [25,26].

Third, at the amino acid level, a versatile solution has to be sought rather than a specific one. In fact, even for identical secondary structures, the geometry (triple helix, α-coiled, β-sandwich…) and/or the symmetry of the protein interfaces also affect the patterns at the amino acid levels [11,17,18,20,27,28,29,30].

For a geometry of interface made of interacting β-strands (β-interfaces), dimers are the main stœchiometry studied, particularly when considering dataset analysis [21,31,32,33,34].

Here, we report the analysis of the β-interfaces of 40 soluble protein oligomers whose stoechiometries are from trimers to octamers. We used our tailor made program Gemini to select hot spots and to produce an interaction network -or a graph- of the subset of interactions that composes an interface [15]. Gemini quantitative and qualitative analyses reveal relatively long β-interfaces enriched with charged residues scattered within the interface. More precisely, arginine residues are preferred at N- and C- terminal extremities whereas histidine and glutamic acid residues are more frequent in the middle of the interfaces. Such a broad charge distribution has never been observed previously in dimeric β-interfaces or in intramolecular β-interactions.

Materials and Methods

Interfaces by Gemini

The computer programs (Gemini) relevant to the present paper have been described previously [15]. In summary, Gemini characterizes an interface as a subset of amino acids in interaction, or "hot spots". They emerge after a purely geometrical analysis of the 3D atomic structure of the protein, well described in the indicated publication. Gemini is equipped with an effective tool (GeminiGraph) that represents interfaces by (bipartite) graphs (Fig. 1). Throughout the paper, the graphs -and so the interfaces- are also referred to as 'interaction networks' or simply as 'networks'. Briefly, the two segments S1 and S2, of an interface are represented by two parallel rows. The interacting amino acids selected by Gemini are indicated by 'X' and the non interacting ones by dots '.' (Fig. 1C). The 'X' amino acids are the hot spots of the interface. The interactions (I) are illustrated by lines connecting two 'X'. The version used here includes the name of the amino acids at positions 'X', following the one-letter code. In few cases, the β-interface is so intimately close to a different interface geometry that Gemini keeps them together in the same interface region (see Table S2 and Dataset S1). In the present work only the β-interface part has been used; the corresponding graphs have therefore been manually annotated (supplementary material).

A supplementary feature has been added to Gemini, which describes the interfaces as two interaction sub-networks. One of them only includes interactions between backbone atoms (BB sub-network), the other interactions with at least one side chain atom (SC sub-network). The interactions of the BB sub-network (I_{BB}) are represented with dashed lines whereas those of the SC sub-network (I_{SC}) are represented with solid lines. X_{SC} and X_{BB} are the side chain and backbone hot spots, respectively.

Circular proteins

This is also a new addition to Gemini especially relevant to the present work. The goal of this part of the code is to recognize

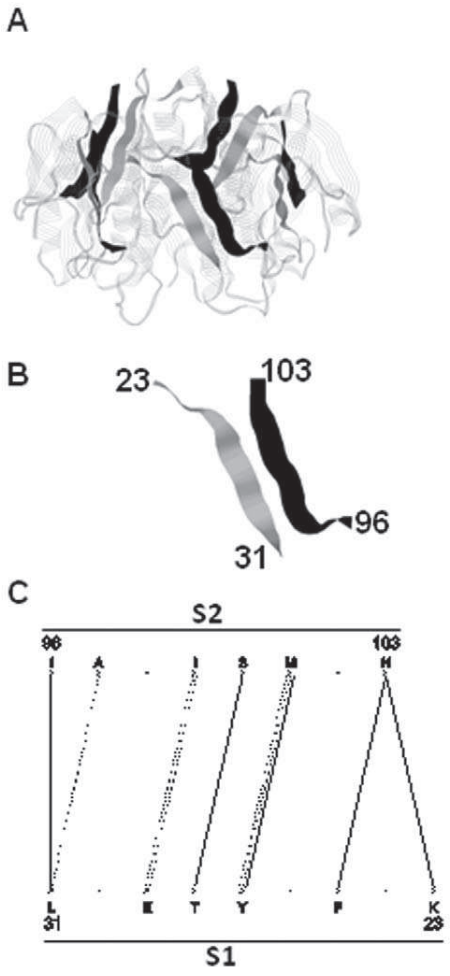


Figure 1. Example of one β-interface geometry. A. The x-ray structure of the whole cholera toxin B pentamer ($CtxB_5$) is shown in strands (PDB code: 1EEI) [66]. The two strands of the β-interface are highlighted in black and grey in ribbons. The image has been generated using Rasmol. **B.** The β-interface is made of the association of the segment composed of amino acids 23 to 31 on one chain (segment 1) and of the segment composed of the amino acids 96 to 103 on the adjacent chain (segment 2). **C.** Gemini graph of the CtxB β-interface. S1 and S2 stand for segments 1 and 2.
doi:10.1371/journal.pone.0032558.g001

circular homo-oligomers (oligomers made of the same protein chain). The program classifies proteins into two classes: circular homo-oligomers and the rest that can contain hetero-oligomers and non circular homo-oligomers. For short, we call it non-circular (NC). The input information is the three-dimensional structure of PDB. No other database or author's annotation is used. The first step in the classification recognizes as NC those proteins whose chains are composed of different numbers of residues. Actually, given that in PDB files there can be additional or missing residues, an error of 25% is tolerated on the differences in the number of residues. The remaining proteins are therefore good candidates to be homo-oligomeric. In a second step, the program tries to find the first amino acid common to all the subunits. From it, five other common amino acids must be found, located at 15%, 30% and so on, of the sequence. If this step fails, the protein is NC. If it succeeds, the protein is very likely to be a homo-oligomer so a third step is needed to evaluate the spatial organization of the subunits. This is simply done by comparing the

distances of the Cα of the six common amino acids already found. If the protein is a circular n-oligomer, there must be n identical distances (a tolerance of 5 Angstrom is used) otherwise the protein is NC. This algorithm is effective in finding circular homo-oligomers but is not enough to fully discriminate within the NC class. There are some false negatives, namely proteins that are circular homo-oligomers but are recognized as NC. This has the only effect of slightly reducing the size of our dataset. We did not observe false positives.

Cytoscape (http://www.cytoscape.org/)

It is an open source bioinformatics software platform for visualizing molecular interaction networks and biological pathways and integrating these networks with annotations, gene expression profiles and other data. Although Cytoscape was originally designed for biological research, now it is a general platform for complex network analysis and visualization. Among the several types of interaction data supported, the format SIF (simple interaction format) was used for the present paper.

RING (Residue Interaction Network Generator)

It is a web server with software for transforming a protein structure (in PDB format) into a network of interactions. Nodes represent single amino acids in the protein structure, while the edges represent the non-covalent bonding interactions that exist between them [35,36,37]. The interaction network and the edge attributes are stored in files with the SIF format. These files can then be easily loaded into CYTOSCAPE to visualize and manipulate the network [35,36,37]. In the present study, RING and CYTOSCAPE were used to produce and visualize the network of hydrogen bonds for the proteins of the dataset.

Statistics

Median, quartile- The median is the value that splits the dataset into two equally populated subsets (above and below the median). For example, for 40 cases and a median of 180 amino acids in size, there are 50% of the cases with a length above 180 and 50% with a length below 180 amino acids. The quartile is the value at which the dataset it divided into four parts, equally populated with the 25% of the samples. The lower separation point is the first quartile, the middle one is the median and the higher is the third quartile.

Global and Local propensity

The ratio between the amino acid frequency in a domain and the amino acid frequency in a database is called "*global propensity*". If the global propensity is above 1, the amino acid is "preferred" in the domain and if the propensity is below 1, the amino acid is "disfavored" in the domain. The "*local propensity*" is defined by the ratio between the amino acid frequency in a particular position (e.g. corner) of a sub-domain (e.g. β-interface) and its frequency in all the other positions in the sub-domain. A local propensity above 1 means the amino acid is preferred in that position than anywhere else in the sub-domain [38]. On the contrary, a local propensity below 1 means the amino acid is disfavored in that position compared to elsewhere in the sub-domain. The corner positions are the amino acids located at the four outer positions on a segment: two outer positions on each side of the segment. So each segment has four amino acids positioned on corners and two outer interactions. The central positions are anywhere else on the segment.

Secondary-structure prediction

GOR IV software was used to perform the secondary structure prediction of the segments of the proteins of the dataset. The secondary structure of each segment of the dataset was predicted (40×2 cases) considering all the wild-type amino acids of the segments and not only the -X-. Then, a residue was mutated and the secondary structure prediction was performed again. When a mutation affected the wild-type original secondary structure prediction, the mutated residue was considered important for the secondary structure of the segment. Hydrophobic residues of the BB or of the SC sub-networks, centrally located or at corners were mutated to charged residues (e.g. K, D, R, E, H). If one of the mutations affected the secondary structure prediction, mutation to other charged amino acids was not essayed. Polar and charged residues of the BB sub-networks centrally located in the full network, were also mutated to either polar or hydrophobic residues.

Probability

Let's call p_c the probability to find in an interface, a charged amino acid. We now evaluate p_{cc}, the probability to have at least one charged amino acid in (at least) one of the corners. This is evaluated as follows:

$$p_{cc} = 4 * p_c * (1-p_c)^3 + 6 * p_c^2 * (1-p_c)^2 + 4 * p_c^3 * (1-p_c) + p_c^4 = 1-(1-p_c)^4$$

where each addendum is respectively the probability to find: a charged amino acid in one corner only, a charged amino acid in two corners, a charger amino acid in three corners, a charged amino acid in all corners. Everything holds true for the corner probability within one of the sub-networks, provided p_c is the corresponding probability.

Reagents and buffers

Cholera toxin B pentamer ($CtxB_5$) and all other chemicals were obtained from Sigma. McIlvaine buffer (0.2 M disodium hydrogen phosphate, 0.1 M citric acid, pH 7.0), PBS and 0.1 M KCl/HCl at pH 1.0 were used. All buffers were filtered through sterile 0.22 μm filter before use. Synthetic peptides were ordered from proteogenix (www.proteogenix.fr).

SDS-PAGE analysis

SDS-PAGE (15% or 12%) were performed with a Bio-Rad mini-Protean 3 system using the Laemli method [39]. The gels were stained with Coomassie blue. 1 μg of sample was loaded on each lane of the gel.

Reassembly of CtxB into native pentamer

The conditions used for reassembly were adapted from elsewhere [40]. Briefly, native $CxtB_5$ was acidified in 0.1 M HCl/KCl at pH 1.0 for 15 min at a final toxin concentration of 86 μM, to induce the toxin dissociation into monomers (MW~11 600 kDa). The toxin was subsequently diluted to a final concentration of 8,6 μM, in McIlVaine buffers at pH 7.0 to promote reassembly. The samples were incubated for 15 min at 23°C before analysis by SDS-PAGE. The reassembly into native CtxB pentamer was inferred from SDS-PAGE analyses since $CtxB_5$ is stable in SDS-containing buffers and migrates in a gel, run on ice,with an apparent molecular weight characteristic of the B-subunit pentamer (MW~55 000 kDa). Only the native pentamer is SDS-resistant. The CtxB concentration for all experiments refers to the monomeric concentration.

Reassembly of CtxB in presence of peptides

The toxin reassembly was measured in presence of synthetic peptides whose sequences correspond to the toxin β-interfaces

sequences (segments 1 and 2). The peptides were added in the neutralizing buffer at a molar ratio peptide to protein of 20. The reassembly conditions were identical to the one used for the toxin alone.

Results

The primary goal of the analysis is to seek protein interface features within a dataset of protein oligomers sharing only a common geometry of interfaces. This is inspired by the success obtained for α-coiled interfaces [17,18,19]. The second objective is to see if the features can be rationalized in term of assembly mechanisms. The interfaces are analyzed using our tailor made program Gemini, which considers interfaces as interaction networks and allows both quantitative and qualitative studies [15].

The dataset

The dataset was built by screening the Protein DataBank (PDB) [41]. First, cyclic protein oligomers were selected so all the cases had identical symmetry (circular, C_n). To this purpose a program called "Circular" (materials and methods) was made. In total 502 protein oligomers were identified with stœchiometries from 3 (trimer) to 8 (octamer) (Table 1). Stœchiometries above 8 contained too few cases to be considered. Second, the secondary structure of the protein interface was chosen as two interacting β-strands at least 4 amino acids apart on the individual chain. The two interacting β-strands had to be different in their amino acid sequences (Fig. 1). Each strand is called a segment. Segment 1 (S1) appears first (N-terminal side) followed by segment 2 (S2) (C-terminal side) on the primary sequence. This geometry is referred to as a β-interface throughout the paper. Third, dimers, hetero-oligomers, transient oligomers, viral and membrane proteins were discarded from the dataset as their interfaces are likely to be differently programmed. After selection, the dataset was made of 40 protein interfaces but the list is non exhaustive.

Properties of the whole chain proteins of the dataset

The protein oligomers are produced by organisms from the three super-kingdoms of life with 2% of archea, 75% of bacteria and 23% of eukaryotes (Table S1). For comparison, there are 8%, 54% and 38% of archea, bacteria and eukaryotic protein oligomers for the stœchiometries from 3 to 8 in the PDB. The atomic structures (PDB) of the protein oligomers of the dataset are shown in figure 2 to illustrate the diversity of their quaternary, tertiary (folds) and secondary structures. The folds are also represented by the SCOP superfamily codes in Table S1 [42].

The secondary structure content of the whole chains is also extensively variable with on average on the dataset 30±20; 40±20 and 30±10% of α-, β- and random coiled structures. This is illustrated in figure 3 with the structures of the chaperone 1Q3S and of the oxidoreductase 1PVN which have a high content of α-structures (60 and 46%, respectively).

The distribution of the whole chain lengths is broad as can be seen on the histogram on figure 4. The median length is 160 amino acids for an interquartile of 148 amino acids. The average length is 203±127 amino acids, value slightly smaller than the average length of monomeric proteins (~300 amino acids) (Tables S1) [1]. This might be due to the measurement of the protein lengths from the PDB sequences which contain gaps due to crystallization or diffraction issues.

The circular trimers are the most represented (67%) against an average of 7±4% for the other stœchiometries (Table 1). The abundance of trimers might be related to the fact that the PDB over-represents low stœchiometries, dimer and trimer in particular, owing to the difficulties in crystallization. The β-interface geometry represents on average 8% of the circular protein oligomers (40/502) in good agreement with a previous measurement in dimers [21].

In summary, the protein oligomers of the dataset are produced by diverse organisms and cover a variety of functions, folds, amino acid lengths and stoichiometries (Table S1). Not surprisingly, the alignment of their amino acid sequences has no worthy of notice homology (not shown). Hence the dataset is characterized by a large heterogeneity.

Global beta interface characteristics

Gemini's interaction networks (or graphs) of the β-interfaces are in Dataset S1. The length and the number of hot spots (-X-) of each β-interface, are determined using the Gemini graphs (materials and methods). Both are counted considering the two segments, S1 and S2, of the interface (Table S2). The statistics on hot spots, interface length and number of interactions are summarized in Table 2. The average length and number of hot spots for the segment S1 or for the segment S2, are similar, indicative of indistinguishable characteristics of the two β-strands of the β-interfaces. The number of interactions between two hot spots (X) involved in the β-interfaces (I_β) is also provided by Gemini (Table S2 and Table 2).

The length, the hot spot number and the interaction number (I_β) have medians and interquartile ranges fairly similar to their respective average and standard deviation values indicative of a relative homogeneity of these features throughout the dataset (Table 2). Yet there is no visible common topological feature within the graphs of the β-interfaces or any specific chemical composition compared to the whole chains (Table 3). A slightly different chemical composition appears when the hot spots are considered instead of all the amino acids of the two segments S1 and S2 (Table 3). No particular sequence homology was observed upon alignments of the S1 and S2 segments (not shown).

It was then assumed that common features might be somehow diluted in a 'background' noise.

As the backbone atoms are identical for the twenty amino acids, it was possible that counting them in the chemical properties of the β-interfaces 'hid' some chemical specificity only distinguishable on the side chain atoms. Likewise, only the backbone atoms might

Table 1. Circular protein oligomers containing a β-interface.

Category	Trimer	Tetramer	Pentamer	Hexamer	Heptamer	Octamer	Total
Circular oligomers	339	39	54	43	22	5	502
β-interface	13	6	11	4	4	2	40
Circular oligomers (%)	67 (339/502)	8	11	9	4	1	100

doi:10.1371/journal.pone.0032558.t001

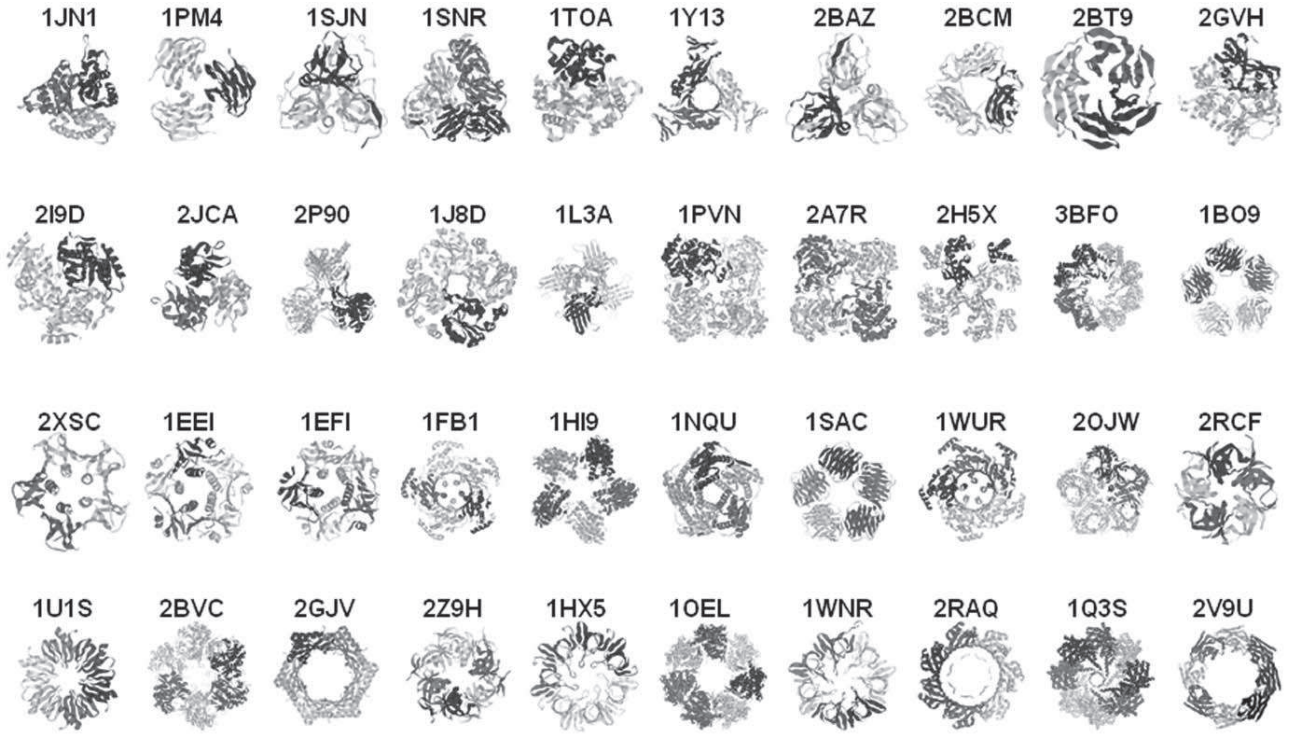


Figure 2. x-ray structures of the protein oligomers of the dataset. The respective PDB codes are indicated above the structures. The figure was made using RasMol. Each chain is shown in a different color.
doi:10.1371/journal.pone.0032558.g002

carry topological information. Moreover, previous studies on protein interfaces had indicated the importance of distinguishing main chain (backbone atoms) contacts from side chain contacts [2,43,44].

Accordingly, the graphs of the β-interfaces were partitioned in two sub-graphs, one made of the backbone interactions (one atom of the backbone per segment, BB sub-networks) and one made of the side chain interactions (one atom of the side chain per segment or one atom of the side-chain on a segment and one atom of the backbone on the other segment, SC sub-network). They are shown in supplementary material 1 (Dataset S1). The interactions within the BB sub-networks are illustrated with dashed lines whereas the interactions within the SC sub-networks are illustrated with solid lines (see also materials and methods). It is important to note that the BB and SC sub-graphs can be considered individually (not considering the whole graphs) or within the whole graph. This nuance is important and when the two sub-networks are considered together, we will refer to as the "full" graph or the full network.

Characteristics of the BB sub-networks

The discrimination of the BB and SC sub-networks revealed significant features shared by the β-interfaces.

The BB sub-networks appeared characterized by common topological features but not by chemical specificities. First, different patterns of interactions show up in the BB sub-graphs. The first one, which appears in 19 graphs, is referred to as the "ladder" pattern because the BB interactions are running parallel to one another (Fig. 5). The second pattern which appears in 8 graphs is referred to as the "V-shape" pattern because it's a triplet interaction in the shape of a -V- (Fig. 6). The patterns are defined by elementary interaction blocks. One block "X.X" on one segment interacts with one block "X.X" on the other segment in the ladder pattern. One "X" on one segment interacts with one block "X.X" on the other segment in the V-shape pattern. The elementary blocks appear singly or in multiple copies. Single versions of the ladder pattern appear in 1PVN, 2OJW, 1U1S and 1HX5 and in multiple copies in 1PM4, 1SNR, 1HI9, 1WUR, 2BCM, 2RCF, 2GJV, 2GVH, 2P90, 1J8D, 1WNR, 2RAQ, 1EEI and 1EFI . There are slightly altered versions of the ladder pattern. One graph (1FB1) is made of one block "X.X" on one segment interacting with one block "X . . X" on the other segment. Two graphs (2I9D and 2RCF) have one block of "XX" on one segment interacting with one block "XX" on the other segment.

Single version of the V-shape pattern can be observed in 2A7R and 2V9U and in multiple copies in 1SJN, 2BAZ, 1L3A, 1NQU, 1OEL and, 1Q3S.

There are also 5 graphs made of a mix of ladder and V-shape patterns (1Y13, 2I9D, 2H5X, 3BFO, 2Z9H).

The second topological information of the BB sub-networks is the fact that the ladder and the V-shape patterns appear related to the arrangement of the secondary structures of the β-interfaces. Indeed, they are observed mostly in anti-parallel and in parallel intermolecular β-strand interactions, respectively, and the pattern shapes' are reminiscent of the anti-parallel and parallel intramolecular main chain hydrogen bond networks found in β-sheets (Figs. 5B & 5C and 6B & 6C). To determine whether Gemini's BB networks were related to intermolecular hydrogen bonds, the program RING (materials and methods) was used, showing that out of the 100 atoms detected by RING as participating in hydrogen bonds, 98 are Gemini's backbone atoms. This is likely due to the selection process of Gemini which retains the closest atoms [15]. Gemini detects slightly more backbone atoms and bonds than RING (139 against 100) due to the fact that Gemini is

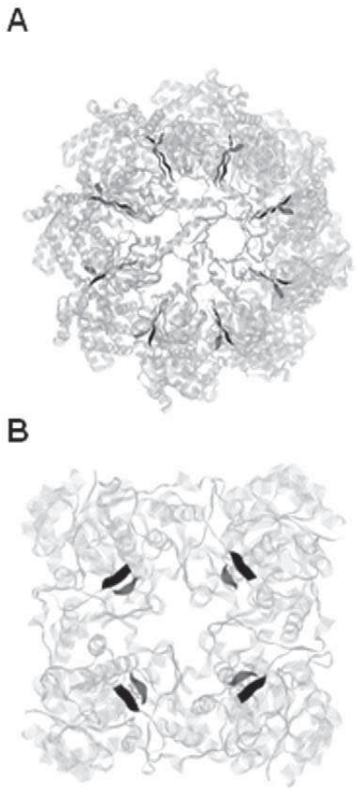


Figure 3. Protein oligomers containing a β-interface. A. The 1Q3S octameric bacterial chaperone [67] and **B.** The 1PVN tetrameric protozoa oxidoreductase [68]. Both structures are represented using RasMol. The chains are colored in light grey and the secondary-structures are represented by helices and strands. The β-strands of the interfaces are colored in black and dark grey in ribbons.
doi:10.1371/journal.pone.0032558.g003

able to detect the double interactions per amino acids observed in the hydrogen bond network of intramolecular β-sheets (Fig. 5B & 5C and 6B & 6C). Thus, the BB sub-networks describe intermolecular β-sheets. This is confirmed by the observation that the graphs which have no BB interaction (1JN1, 1T0A, 2JCA, 1B09, 2XSC, 1SAC) or only one BB interaction (2BT9 and 2BVC)

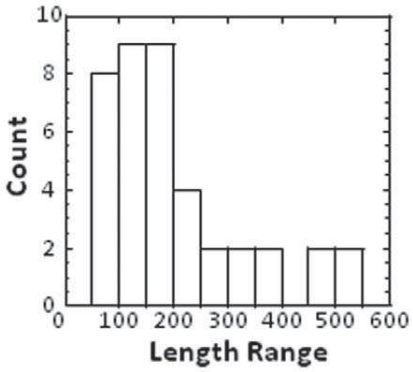


Figure 4. Histogram of the whole chain lengths. The length of the whole chain (range) is indicated on the x axis as the total number of amino acids.
doi:10.1371/journal.pone.0032558.g004

Table 2. Statistics on the lengths of the dataset.

Sample	Average	SD[a]	Median	Q3-Q1	Q3[b]	Q1[b]
Length	17	6	17	7	19	12
Hot spot 'X'	12	4	12	5	14	9
Iβ	10	4	10	5	12	7

[a]SD stands for standard deviation.
[b]Q stands for quartile. The statistics are defined in materials and methods.
doi:10.1371/journal.pone.0032558.t002

are not intermolecular β-sheets but are two rather perpendicular interacting β-strands, as can be seen on their respective PDB.

The BB sub-networks (X_{BB}) cannot be distinguished from the whole chains by a specific chemical composition (charged, polar and hydrophobic amino acids). Yet, they are dominated by hydrophobic properties: half of the amino acids of the BB sub-networks are hydrophobic and a third of the interactions are purely hydrophobic (Table 3 and table 4).

The global propensity (materials and methods) of the hydrophobic amino acids of the BB sub-networks was measured to evaluate which hydrophobic amino acids were over-represented in the β-interfaces compared to the whole chains (Table 5). A global propensity above 1 indicates a hydrophobic amino acid "preferred" in the BB sub-networks and on the contrary, a global propensity below 1, indicates a hydrophobic amino acid depleted in the BB sub-networks. Methionine (M), cysteine (C), tryptophane (W), isoleucine (I) and valine (V) are preferred in the BB sub-networks whereas proline (P), alanine (A), glycine (G) and leucine (L) residues are not favored in the BB sub-networks. The phenylalanine is equally present in the BB sub-networks and in the whole chains of the dataset (Global propensity around 1).

Characteristics of the SC sub-networks

In contrast to the BB sub-networks, the SC sub-networks have no topological information but some chemical specificity. In fact the SC sub-networks present an average chemical composition significantly different from the whole chains with a decrease of the percentage of hydrophobic amino acids in favor of an increase of the percentage of charged amino acids (Table 3). The percentage of polar residues remains similar for the SC sub-networks and the whole chains. This observation is even more obvious when the interactions (I_{SC}) are considered instead of the individual amino acids (X_{SC}), as the SC sub-networks have 5 times more purely charged interactions (Ch-Ch) than the BB sub-networks (Table 4). The SC sub-networks also have twice less purely hydrophobic interactions (F-F) than the BB sub-networks (Table 4).

Table 3. Average chemical composition, in percentage, of the amino acids of the whole chain of the protein dataset, of the two segments of the interface S1+S2) and of the hot spots of S1 and S2. SC and BB stand for side chain and backbone amino acids, respectively.

Interfaces	whole	S1+S2	S1+S2 'X'	X_{SC}	X_{BB}
Charged	24±17	24±10	28±14	30±17	23±16
Polar	23±15	26±14	29±16	29±17	27±24
Hydrophobic	53±34	50±12	45±15	41±15	50±14

doi:10.1371/journal.pone.0032558.t003

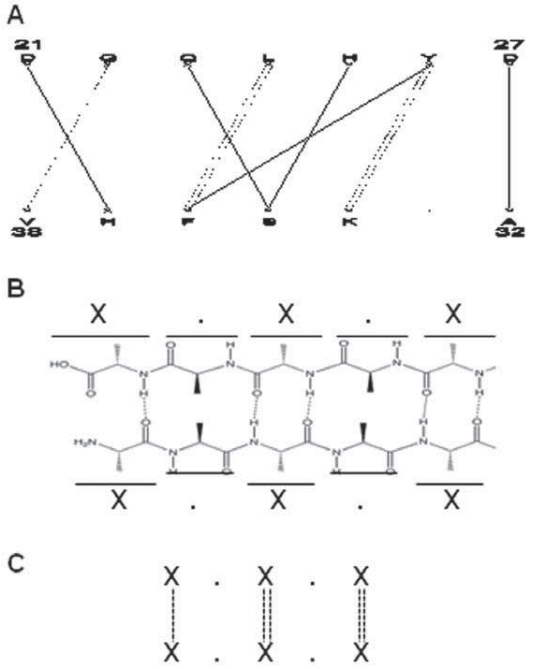


Figure 5. Anti-parallel BB sub-network and intramolecular hydrogen bond network. A. Gemini graph of an anti-parallel intermolecular β-interface **B.** Schematics of the hydrogen bond network of anti-parallel intramolecular β-sheet. **C.** Ladder pattern observed in BB sub-network and also visible in anti-parallel intramolecular β-sheet.
doi:10.1371/journal.pone.0032558.g005

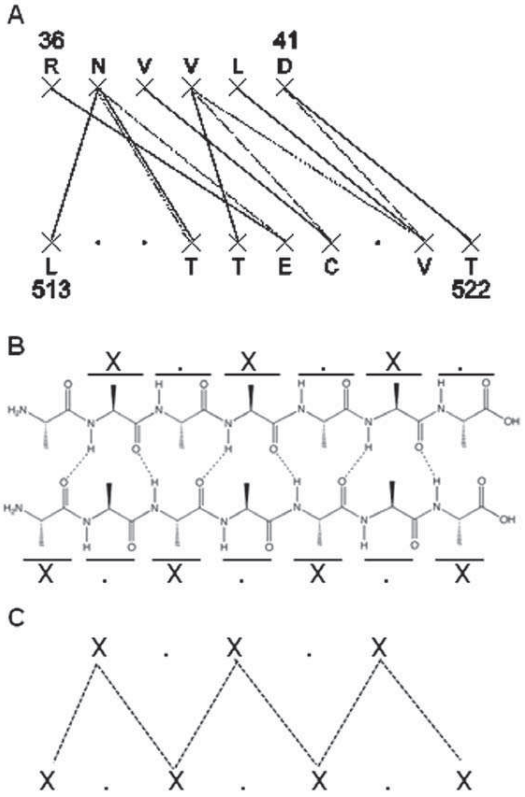


Figure 6. Parallel BB sub-network and intramolecular hydrogen bond network. A. Gemini graph of a parallel intermolecular β-interface **B.** Schematics of the hydrogen bond network of parallel intramolecular β-sheet. **C.** Ladder pattern observed in BB sub-network and also visible in parallel intramolecular β-sheet.
doi:10.1371/journal.pone.0032558.g006

The global propensity (materials and methods) of the charged residues of the SC sub-networks compared to the whole chains is reported in Table 6. A charged amino acid with a global propensity above 1 is "preferred" in the SC sub-networks whereas a charged amino acid with a propensity below 1 is depleted. Apart from the histidine, which has a global propensity slightly above 1.0, all the charged residues of the SC sub-networks have a global propensity around 1.

The local propensity of the charged amino acids in the SC sub-networks was analyzed considering corner (the four outer SC amino acids) and central (non corner) positions (Table 7 and table 8, respectively). The local propensity (material and methods) is the ratio of the frequency of an amino acid in a particular position (e.g. corner) within a local structure (e.g. the β-interfaces) and of the frequency of the same amino acid in any other position within that local structure [38]. There are almost as much charged amino acids at corners than at central positions (44% in corner positions). But the two positions are made of different types of charged residues. Arginine (R) residues are more frequent at corners (local propensity above 1 in table 7) whereas it is glutamic acid and histidine residues which are favored centrally (local propensity above 1 in table 8). The lysine and aspartic acid residues have no local preferences (local propensity around 1 in both table 7 and table 8).

Comparison of BB and SC sub-networks

There exist several differences between the BB and the SC sub-networks (Table S3). There are 6 ± 3 I_{SC} interactions for only 4 ± 2 I_{BB} interactions. Additionally, there are 9 ± 4 X_{SC} amino acids for only 5 ± 3 X_{BB} amino acids. An amino acid with one atom involved in a BB interaction and one atom involved in a SC interaction is counted twice, one per network. But an amino acid having several atoms participating to the same network is counted only once. Thus, on average, the SC sub-network is bigger than the BB sub-network with roughly 60% of the interface amino acids and interactions devoted to it.

When considering the full graphs, it appears that the BB sub-networks are depleted of interactions and of hot spots at corners having only two graphs with two I_{BB} in the outer positions (1NQU and 2Z9H) and only 11 with one I_{BB} in the outer position (1Y13, 2BCM, 1PVN, 2A7R, 2H5X, 3BFO, 1EFI, 2OJW, 1U1S, 1WNR AND 1Q3S). In contrast, 28 graphs have two SC interactions in the outer positions and 39 (out of 40) have at least one. Likewise, the SC sub-networks are depleted of interactions and of hot spots

Table 4. Chemical composition of the interactions (amino acid -i- of segment 1 with amino acid -j- of segment 2 or vice-versa, data are added together) in the SC and in the BB (bracket) networks of the β-interfaces.

Chemical properties	Charged	Polar	Hydrophobic
Charged	17% (3,5%)		
Polar	13% (13%)	10% (9%)	
Hydrophobic	18% (23%)	25% (21%)	16% (30%)

doi:10.1371/journal.pone.0032558.t004

Table 5. Global propensity of the hydrophobic residue in the BB sub-networks.

Hydrophobic	Number in the BB sub-networks	Percentage in the BB sub-networks	Number in the Whole chains	Percentage in the whole chains	Global propensity
I	26	0.20	577	0.13	1.6
L	17	0.13	700	0.16	0.8
V	27	0.21	714	0.16	1.3
A	13	0.10	756	0.17	0.6
C	4	0.03	95	0.02	1.5
M	11	0.09	180	0.04	2.1
F	9	0.07	295	0.07	1.1
G	15	0.12	714	0.16	0.7
P	4	0.03	351	0.08	0.4
W	3	0.02	82	0.02	1.3
Total	129		4464		

doi:10.1371/journal.pone.0032558.t005

at central positions. There are 86 I_{SC} centrally located for a total of 240 I_{SC} (36%) and 143 X_{SC} centrally located for a total of 374 X_{SC} (38%). In the BB sub-networks, there are 86 I_{BB} centrally located for a total of 156 I_{BB} (55%) and 131 X_{BB} centrally located for a total of 219 X_{BB} (60%). This means that in a typical arrangement, the SC sub-network spatially contains and surrounds the BB one.

Consequently, the corners of the SC sub-networks are enriched with charged residues (32 graphs out of 40, 80%) while those of the BB sub-networks are depleted (10 graphs out 34: 29%). Similarly, the BB sub-networks are enriched centrally with hydrophobic residues (72 central hydrophobic residues for 110 in total: 65%) while the SC sub-networks are depleted (41 central hydrophobic residues for 101 in total: 41%).

Hence, the relative position of the sub-networks provides enrichment (or depletion) of a chemical property without having to vary the absolute number of amino acids of that property in the sub-networks. For example, there are 110 and 101 hydrophobic residues in the BB and SC sub-networks, respectively. Also, the probabilities of finding a charged residue in the corner of the SC or of the BB sub-networks, based on their respective chemical properties (Table 3), are indeed very similar 76% and 65%, respectively (materials and methods). Yet by positioning the X_{BB} centrally, the charged X_{SC} appear more frequently at corners.

Rationalization of the BB and SC features

Once common features are identified within the β-interfaces of the dataset, the next question is: can those features be rationalized in term of protein assembly or interface formation?

The first argument in that direction, is the weight of the β-interactions (Table S2). I_{β} are the interactions involved in the β-interface region of the protein oligomers of the dataset. Now, the total number of intermolecular interactions (I_{tot}) in a whole chain is the number of interactions in all the interface regions. I_{tot} is provided by Gemini. The average number of intermolecular interactions (I_{av}) per chain is the total number of interactions (I_{tot}) divided by the number of interface regions. The weight of the β-interactions is measured by the ratio -I_{β}/I_{av}- which gives the amount of interactions in a β-interface compared to the average number of interactions in the whole chains. On average, there are twice more interactions in the β-interfaces than in the whole interface (1.8 ± 0.6). The high number of interactions due the beta geometry is consistent with a role of the β-interfaces in the assembly mechanism.

The data indicate that the BB sub-networks are related to the secondary structures of the interfaces and that they are enriched in hydrophobic residues and hydrophobic interactions. In order to test the involvement of the hydrophobic residues in the secondary structure of the interface, the effect of their mutation on secondary structure prediction was investigated.

The secondary structure of the segments (S1 and S2) with the wild-type (WT) sequence was predicted using GOR IV and compared to the prediction of the same segment after a point mutation of one hydrophobic residue. The mutation of centrally located hydrophobic residues to a charged residue (e.g. K, D, R, E, H) altered the secondary-structure prediction in 83% of the cases. The mutation of hydrophobic residues located at corners to

Table 6. Global propensity of the charged residue in the SC sub-networks.

Charged	Number in the SC sub-networks	Percentage in the SC sub-networks	Number in the whole chains	Percentage in the whole chains	Global propensity
R	19	0.16	403	0.18	0.9
E	31	0.27	584	0.27	1.0
K	26	0.22	497	0.23	1.0
D	24	0.21	507	0.24	0.9
H	13	0.11	188	0.09	1.3
Total	113		2179		

doi:10.1371/journal.pone.0032558.t006

Table 7. Local propensity of the corner charged residue in the SC sub-networks.

Charged	Number in the corner position	Percentage in the corner position	Number in the SC sub-networks	Percentage in the SC sub-networks	Local propensity
R	12	0.24	19	0.17	1.4
E	11	0.22	31	0.27	0.8
K	12	0.24	26	0.23	1.0
D	11	0.22	24	0.21	1.0
H	4	0.08	13	0.12	0.7
Total	51		113		

doi:10.1371/journal.pone.0032558.t007

charged residue, also disturbed the secondary-structure prediction but to a much lesser extent (44% of the cases). In the same way, the mutation of polar or of charged residues of the BB sub-networks centrally located, to hydrophobic, charge or polar amino acids affected the secondary-structure prediction in only 44% of the cases.

We then measured the local propensity of the hydrophobic residues located centrally in the BB sub-networks and affecting the 2D structure prediction (Table 9). It appears that among the secondary-influencing hydrophobic residues centrally located, the valine (V) and the phenylalanine (F) are preferred (local propensity above 1). The leucine (L), the isoleucine and the methionine (M) appear neutral in the central position (local propensity around 1). Tryptophan (W), proline (P), glycine (G), alanine (A) and cysteine (C) are not favored (local propensity below 1).

The local propensity results were tested using secondary-structure prediction again. Mutations of central hydrophobic amino acids of the BB sub-networks to hydrophobic amino acids which have a local propensity above 1 were expected to have a secondary-structure prediction identical to the wild-type one. This is referred to as the amino acid having a positive versatility (act as wild-type amino acid). On the contrary, mutations to amino acid with a local propensity below 1 were expected to alter the wild-type secondary-structure prediction. These amino acids are referred to as having a negative versatility. In total 331 mutations-predictions have been performed and on average 69% behave as expected (229/331). Both the versatilities are giving similar results with 67% (116/172) of the mutations to amino acids of positive versatility not affecting the secondary structure prediction and 71% (113/159) of the mutations to amino acids of negative versatility affecting it.

This is consistent with the involvement of the features of the BB sub-networks in the secondary structure formation of the β-interfaces.

The SC sub-networks have no topological information and therefore cannot be related to geometrical features. But they have enrichment in charged residues and more precisely a specific distribution of the type of charges along the interface. This suggests a chemical role of the SC sub-networks in the formation of the β-interfaces, via electrostatic interactions.

We have seen that the local positions of the hydrophobic and of the charged residues of the BB and SC sub-networks were connected to the relative position of the two sub-networks. Now, remarkably for the 11 graphs which have one outer BB interaction, 7 have one charged BB residue at a corner. Following the same drift, the graphs with a low content of SC interactions but made of a majority of BB interactions have a charged BB residue in a corner in 44% of the case (7 out 16 graphs) whereas this occurs only in 12% of the graphs made of a minority of BB interactions (3/24).

So even if having a charged residue in a corner appears a trademark of the SC sub-networks, a corner charged residue is maintained via the BB sub-networks if necessary. This looks like a compensatory or a substitutive mechanism.

A similar phenomenon can be observed for the hydrophobic property of the graphs. On average twice more SC hydrophobic residues are located centrally (1,1 central SC hydrophobic) in graphs made of a minority of BB interactions than in graphs made of a majority of BB interactions (0,45 central SC hydrophobic). More precisely, the number of centrally located hydrophobic residues is maintained at a value of $2{,}8 \pm 0{,}6$ across the dataset with $2{,}2 \pm 0{,}5$ of them affecting the secondary structure predictions (Fig. 7). This value is kept constant using either BB or SC residues, or a balance of both. The mutation of the centrally located hydrophobic residues of the SC sub-networks to charged residue affects the secondary prediction in 83% of the case, as for the BB sub-networks. Thus the regulation of the secondary structure

Table 8. Local propensity of the corner charged residue in the SC sub-networks.

Charged	Number in a NOT corner position	Percentage in a NOT corner position	Number in the SC sub-networks	Percentage in the SC sub-networks	Local propensity
R	7	0.11	19	0.17	0.7
E	20	0.32	31	0.27	1.2
K	14	0.22	26	0.23	1.0
D	13	0.21	24	0.21	1.0
H	9	0.14	13	0.12	1.2
Total	64		113		

doi:10.1371/journal.pone.0032558.t008

Table 9. Local propensity of the central hydrophobic residue of the BB sub-networks affecting the 2D-structure prediction.

Hydrophobic	Number in central position	Percentage in the central position	Number in BB sub-networks	Percentage in BB sub-networks	Local propensity
I	11	0.20	26	0.20	1.0
L	8	0.15	17	0.13	1.1
V	18	0.33	27	0.21	1.6
A	3	0.06	13	0.10	0.6
C	1	0.02	4	0.03	0.6
M	4	0.07	11	0.09	0.9
F	5	0.09	9	0.07	1.3
G	3	0.06	15	0.12	0.5
P	1	0.02	4	0.03	0.6
W	0	0.00	3	0.02	0.0
Total	54		129		

doi:10.1371/journal.pone.0032558.t009

through hydrophobic amino acids located centrally is organized by the BB sub-networks in most cases. But the BB sub-networks can be substituted by the SC sub-networks as an alternative.

Such compensatory or substitutive phenomenon is also in favor of the features being involved in the formation of the interface.

No distinction between the stœchiometries was found for any of the properties of the β-interfaces (not shown).

Autonomous β-interface segments

As mentioned earlier, the features describing the β-interfaces are rather homogeneous compared to the heterogeneity observed for their whole chains. In addition, it seems possible to associate the β-interface features to geometrical and chemical properties. This hinted the possibility that the β-interfaces had some autonomous capacity to associate in absence of the whole chain. This was further supported by the narrow distribution of the β-interface lengths and by the absence of proportion between the lengths of the β-interface and the length of their respective whole chain (Fig. 8). To test that possibility, a simple experiment was carried out using the pentamer of the cholera toxin B ($CtxB_5$) as a prototype of the β-interfaces (Fig. 1). Conditions to follow the assembly of the $CtxB_5$ *in vitro* had been established previously and are indicated in material and methods [40]. Briefly, the native toxin (Fig. 9, lane 2) is acidified for 15 min at room temperature (RT) to lead to its dissociation into monomers (Fig. 9, lane 3). Subsequently, it is neutralized for 15 min at RT, time during which the reassembly into pentamer takes place (Fig. 9, lane 4). In subsequent experiments, 9mer (P1) or/and 8mer (P2) synthetic peptides with sequences corresponding to S1 (^{23}KIFSYTESL^{31}) and S2 (^{96}IAAISMAN^{103}), respectively, of the wild-type CtxB β-interface were added to the neutralizing buffer. The amounts of CtxB reassembled into pentamer under the different conditions,

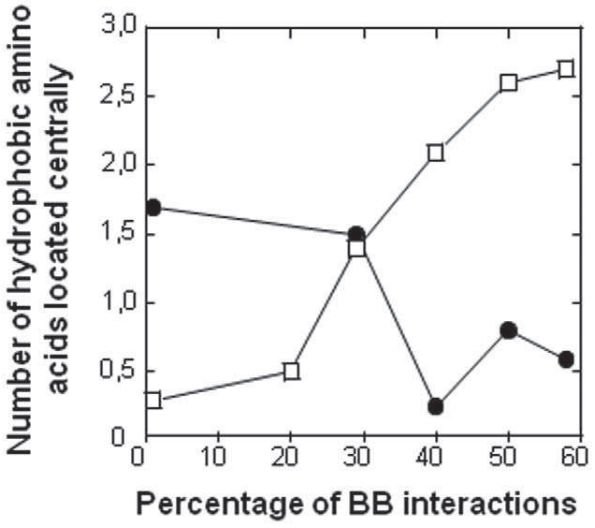


Figure 7. Central hydrophobic residues and percentage of BB interactions. The number of hydrophobic amino acids of the BB (β) or of the SC sub-networks (●) located centrally in the full networks are plotted against the percentage of BB interactions.
doi:10.1371/journal.pone.0032558.g007

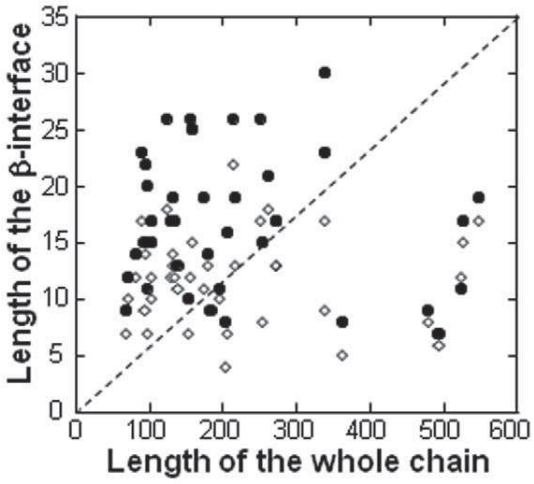


Figure 8. Absence of correlation between the lengths of the whole chains and of the β-interfaces. The length of the β-interface (sum of the amino acids of the two segments) of each protein of the dataset is plotted against the length of its respective whole chain (●, 'all amino acids' and ◇, 'X', respectively). If there was a correlation between the size of the whole chain and the size of its interface or the size of its hot spot numbers, the points would appear on the dashed line.
doi:10.1371/journal.pone.0032558.g008

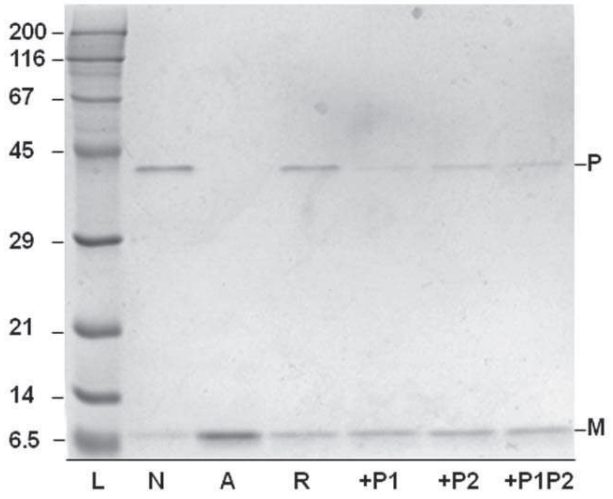


Figure 9. *In vitro* assembly of the cholera toxin B subunit into pentamer ($CtxB_5$). The formation of the CtxB β-interface is monitored by SDS-PAGE. The initial native $CtxB_5$ is indicated in lane 2 (N) whereas the acidified CtxB monomer is indicated in lane 3 (A). The toxin reassembly after 15 min in neutral condition is shown from lane 4 to 7 for the toxin alone (R, lane 4), or with a synthetic peptide of CtxB segment 1 sequence (+P1, lane 5) or with a synthetic peptide whose sequence corresponds to CtxB segment 2 (+P2, lane 6) or with a mixture of both peptides (+P1P2, lane 7). L stands for low molecular weight standard.
doi:10.1371/journal.pone.0032558.g009

were then compared using SDS-PAGE (Fig. 9). The addition of P1 (Fig. 9, lane 5), of P2 (Fig. 9, lane 6) and of P1 and P2 together (Fig. 9, lane 7) strongly inhibited the reassembly of the toxin into pentamer. This indicates that P1 as well as P2 do interfere with the formation of CtxB-CtxB interfaces. P1 inhibited more than P2 and the mixture P1+P2 inhibited more than P2 but less than P1. Thus P1 and P2 must be reacting together.

Discussion

As for the α-coiled interfaces, the choice of a common geometry of interfaces proved to be successful in isolating characteristics among the β-interfaces of otherwise unrelated protein oligomers. The results are thus devoid of potential bias introduced when protein interfaces of proteins with similar folds or similar functions are compared. It was also possible to associate geometrical and chemical properties to the identified features. On one hand, this provides an evaluation of the features so their reliability improves. On the other hand, it also gives some rational about the 'mode of action' of the features in term of interface formation. Thus, using the CtxB model, the role of the hydrophobic and of the charged residues on the formation of the secondary structure and on the formation of the CtxB β-interface, respectively, can be tested. However, the study entirely focuses on the β-interfaces and as such the results are far from providing a full picture of the parameters involved in the assembly of the whole chains of the dataset. As an illustration, we have seen that the mutations of the central hydrophobic residues of the BB sub-networks have little effect on the secondary structure predictions of the whole length sequences (~25%) (not shown). The true essence of the results resides in the observation of interdigitated networks in which the interface features are made through strategic positioning of chemical characteristics rather than through drastic chemical modulation. Thus the search of a sequence of an interface cannot be done as the search of a sequence of a biological function (e.g. active site).

In summary, the β-interfaces are made of two interactions sub-networks. One is involving atoms of the main chain (BB sub-networks) and the other is involving atoms from the side chains (SC sub-networks). The characteristics of the BB sub-networks are related to the hydrophobic residues which seem particularly involved in the secondary structures of the β-interfaces. This is well supported by the fact that the hydrophobic residues favored in the β-interfaces (IVMWC) are also favored in intramolecular β-sheet (IVMCW) [34,45,46,47]. Likewise, the hydrophobic residues disfavored in the β-interfaces (AGP) are disfavored in intramolecular β-sheet (AGP) [34,45,46,47]. There are some discrepancies for the leucine and phenylalanine residues which are favored in intramolecular β-sheets but disfavored or neutral in the β-interfaces, respectively. Intriguingly, these two amino acids are enriched in amyloid β-fiber (LIF) [33]. The role of hydrophobic forces in interfaces (dimers) was previously reported but not in connection with the geometry of the interface [21,48,49] and for review see [2,12,33].

The hydrophobic amino acids of the BB sub-networks are thus devoid of 'intermolecular'specificity since they are shared with intramolecular interactions.

In contrast, the charged amino acids favored in the SC sub-networks present some specificity. First, intra-molecular β-interactions as well as dimeric β-interfaces are rather depleted in charged residues, apart from arginine for the dimeric interfaces ([21,32,33,45,46,50] and for review [2]). On the contrary, in the β-interface side chains, charged residues represent a third of the interfacial amino acids and have only a slight preference for histidine residues. It is interesting that the histidine residue stands out as it is the only amino acid charged under physiological conditions. It is also an amino acid already shown to take part in the assemblies of several protein oligomers [51,52,53]. Second, the β-interfaces of our dataset have an average net charge of -0.5 which differs from the one required for the formation of amyloid β-fiber (net charge of ± 1), another type of β-interface [54,55,56].

The third and most practical information about the charge specificity, resides in the distribution of the charged residues. The arginine residues are frequent at both the corners (N- and C-terminal caps) of the β-interfaces whereas histidine and glutamic acid are favored centrally. Lysine and aspartic acid residues have no preferred position in the β-interfaces.

This is in contrast to parallel intramolecular β-sheet in which positively charged residues (KR) are located at the N-terminal extremities only and negatively charged residues (DE) are present at the C-terminal extremities only [47]. The presence of charges at the N- or C- terminal extremities is believed to act as β-breakers [45,47]. Additionally, the formation of amyloid β-fiber is promoted with positively charged residues (KR) located at the N-terminal extremities of the amyloid β-strands and negatively charged residues (DE) at both the N- or C-terminal extremities [54,55]. Finally, charged residues centrally located are observed in intra-molecular edge β-strands and are thought to prevent their aggregation [34]. Hence, the scattered distribution observed on the β-interfaces differentiates them from other types of intramolecular and intermolecular dimeric β-interactions (Fig. 10).

Altogether the data lead us to propose some hypothesis on the construction mechanism of the β-interfaces following two principles: (i) interfaces are built via geometrical and chemical recognition of the interacting domains and (ii) there are a recognition phase ('binding') and a stabilization phase. The BB sub-networks, via the hydrophobic residues, could provide the geometrical recognition whereas the side chain charged residues could provide the chemical one. It is tempting to speculate that the long arginine residue located at the extremities is employed as a

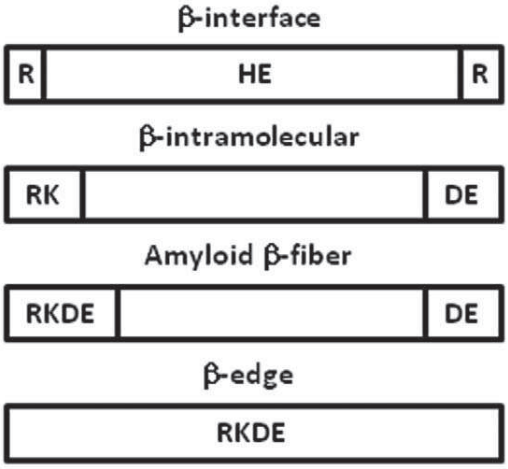


Figure 10. Schematic of the charge distribution in β-interactions. The amino acids are indicated using the single letter code.
doi:10.1371/journal.pone.0032558.g010

hook to promote encounter. The central smaller histidine and glutamic acid residues could act as clips to stabilize the interface. Alternatively, they might, as proposed for the β-edge strands, maintain the two domains soluble prior the recognition.

Some experimental data are consistent with a relation between Gemini's hotspot residues and their involvement in the process of a β-interface formation. For example, the heat labile enterotoxin B (LTB_5) and the cholera toxin B ($CtxB_5$) pentamers, which shares 84% sequence identity and almost superimposable x-ray structures, have nevertheless different assembly mechanisms and different β-interface graphs (1EFI and 1EEI, respectively). The two toxin pentamers have only 14 different amino acids and one of them is in the β-interface (Leu 25 and Phe 25 in 1EFI and 1EEI, respectively). Residue 25 is involved in a I_{BB} in both graphs but leucine and phenylalanine have been measured with different global propensities (Table 5). There are 6 I_{BB} for 4 I_{SC} in LTB_5 compatible with a geometry-regulated assembly as observed experimentally since only folded LTB chains associate [57]. On the other hand, there are 5 I_{BB} for 5 I_{SC} in $CtxB_5$ consistent with a more 'chemically'-regulated assembly also observed experimentally with partially folded CtxB chains capable of associating [40,52]. The presence of a I_{SC} involving a lysine residue only in $CtxB_5$ (K23-N103) also supports a more 'chemically'-regulated assembly. Similarly, shiga-like toxin I and II have different stabilities and different graphs (2XSC and not shown) [58]. In the bacterial hexameric (1U1S) from *Pseudomonas aeruginosa* , the mutation of His 57, to alanine (Ala) or to threonine (Thr) destabilizes the hexamer by disturbing the side chain hydrogen bond network of the His 57 with the side chains of Lys 56 and Ile 59 of the adjacent chain [59]. The His 57 side chain hydrogen bond network is properly seen on the Gemini graph of the β-interface of Hfq (Dataset 1, 1U1S). Disappearance of that network (or changes of that network) for mutant Ala 57 (or for mutant Thr 57) is also seen properly on the Gemini graphs of the mutated Hfq (not shown). Moreover, the conserved main chain hydrogen bond network made of the residues Met 53 and Tyr 55 of chain M with the residues Val 62 and Ser 60 of the adjacent chain is also identified by Gemini (not shown) [60]. However, cautious is necessary with interpreting the graph features. At this stage, they should be used as a tool to formulate hypothesizes for experimental tests.

There are several arguments, mentioned in the result section, supporting the idea that the β-interfaces are independent assembly unit. The most indicative one is the experimental observation that the CtxB β-interface peptides recognize the CtxB individual chains. Such peptides could be called ''assemblons'' by homology to the foldons [61,62]. Some peptides have been found to lead to the trimerization of proteins when genetically added to their sequence, supporting the 'assemblons' concept [63,64,65].

Supporting Information

Dataset S1 Gemini Graphs of the 40 β-interfaces. Each graph appears on a separate page. The stœchiometry and the PDB code of the concerned protein oligomer is indicated on the box in the left hand side of the image. The amino acid number is indicated with the type of amino acid at position X. Segments 1 and 2 appear on two parallel rows. X indicates amino acids involved in atomic interactions according to Gemini. SC and BB interactions are illustrated by solid and dashed lines, respectively [15]. The graphs which interfaces have been annotated manually are indicated with a straight line above the segments. A top (left) and a side view (right) of the x-ray structure of the protein oligomer is shown above its respective graph.
(PDF)

Table S1 Features of the protein oligomers of the dataset.
(DOC)

Table S2 Features of the β-interfaces.
(DOC)

Table S3 Properties of the two sub-graphs.
(DOC)

Author Contributions

Conceived and designed the experiments: CL GF LV. Performed the experiments: MA JZ. Analyzed the data: CL. Contributed reagents/materials/analysis tools: CL GF JZ LV. Wrote the paper: CL.

References

1. Goodsell DS, Olson AJ (2000) Structural symmetry and protein function. Annu Rev Biophys Biomol Struct 29: 105–153.
2. Janin J, Bahadur RP, Chakrabarti P (2008) Protein-protein interaction and quaternary structure. Q Rev Biophys 41: 133–180.
3. Iacovache I, van der Goot FG, Pernot L (2008) Pore formation: An ancient yet complex form of attack. Biochim Biophys Acta.
4. Lesieur C, Vecsey-Semjen B, Abrami L, Fivaz M, Gisou van der Goot F (1997) Membrane insertion: The strategies of toxins (review). Mol Membr Biol 14: 45–64.
5. Kirkitadze MD, Bitan G, Teplow DB (2002) Paradigm shifts in Alzheimer's disease and other neurodegenerative disorders: the emerging role of oligomeric assemblies. J Neurosci Res 69: 567–577.
6. Soto C (2003) Unfolding the role of protein misfolding in neurodegenerative diseases. Nature Reviews Neuroscience 4: 49–60.
7. Klein W, Stine W (2004) Small assemblies of unmodified amyloid [beta]-protein are the proximate neurotoxin in Alzheimer's disease. Neurobiology of aging 25: 569–580.
8. Harrison RS, Sharpe PC, Singh Y, Fairlie DP (2007) Amyloid peptides and proteins in review. Rev Physiol Biochem Pharmacol 159: 1–77.
9. Miller Y, Ma B, Nussinov R (2010) Polymorphism in Alzheimer A amyloid organization reflects conformational selection in a rugged energy landscape. Chemical reviews.
10. Larsen TA, Olson AJ, Goodsell DS (1998) Morphology of protein-protein interfaces. Structure 6: 421–427.
11. Grueninger D, Treiber N, Ziegler MOP, Koetter JWA, Schulze MS, et al. (2008) Designed protein-protein association. Science 319: 206.
12. Tuncbag N, Kar G, Keskin O, Gursoy A, Nussinov R (2009) A survey of available tools and web servers for analysis of protein–protein interactions and interfaces. Briefings in Bioinformatics 10: 217.

13. Cazals F, Proust F, Bahadur RP, Janin J (2006) Revisiting the Voronoi description of protein-protein interfaces. Protein Sci 15: 2082–2092.
14. Shulman-Peleg A, Shatsky M, Nussinov R, Wolfson HJ (2007) Spatial chemical conservation of hot spot interactions in protein-protein complexes. BMC Biol 5: 43.
15. Feverati Ga, Lesieur C (2010) Oligomeric Interfaces under the Lens: Gemini. Plos OneAvailable: http://dx.plos.org/10.1371/journal.pone.0009897.
16. Guidry JJ, Shewmaker F, Maskos K, Landry S, Wittung-Stafshede P (2003) Probing the interface in a human co-chaperonin heptamer: residues disrupting oligomeric unfolded state identified. BMC Biochem 4: 14.
17. Crick FHC (1953) The packing of alpha-helices: simple coiled-coils. Acta Crystallogr 6: 689–697.
18. Lupas A, Van Dyke M, Stock J (1991) Predicting coiled coils from protein sequences. Science 252: 1162–1164.
19. Lupas A (1996) Coiled coils: new structures and new functions. Trends Biochem Sci 21: 375–382.
20. Walshaw J, Woolfson DN (2003) Extended knobs-into-holes packing in classical and complex coiled-coil assemblies. J Struct Biol 144: 349–361.
21. Guharoy M, Chakrabarti P (2007) Secondary structure based analysis and classification of biological interfaces: identification of binding motifs in protein-protein interactions. Bioinformatics 23: 1909–1918.
22. Yan C, Wu F, Jernigan RL, Dobbs D, Honavar V (2008) Characterization of protein-protein interfaces. Protein J 27: 59–70.
23. Davis FP, Sali A (2005) PIBASE: a comprehensive database of structurally defined protein interfaces. Bioinformatics 21: 1901–1907.
24. Tsai CJ, Lin SL, Wolfson HJ, Nussinov R (1996) Protein-protein interfaces: architectures and interactions in protein-protein interfaces and in protein cores. Their similarities and differences. Crit Rev Biochem Mol Biol 31: 127–152.
25. Gao M, Skolnick J (2010) Structural space of protein–protein interfaces is degenerate, close to complete, and highly connected. Proceedings of the National Academy of Sciences 107: 22517.
26. Stein A, Mosca R, Aloy P (2011) Three-dimensional modeling of protein interactions and complexes is going omics. Current opinion in structural biology.
27. Tsai CJ, Lin SL, Wolfson HJ, Nussinov R (1996) A dataset of protein-protein interfaces generated with a sequence-order-independent comparison technique. J Mol Biol 260: 604–620.
28. Grigoryan G, Keating AE (2008) Structural specificity in coiled-coil interactions. Curr Opin Struct Biol 18: 477–483.
29. Calladine CR, Sharff A, Luisi B (2001) How to untwist an alpha-helix: structural principles of an alpha-helical barrel. J Mol Biol 305: 603–618.
30. Hadley EB, Testa OD, Woolfson DN, Gellman SH (2008) Preferred side-chain constellations at antiparallel coiled-coil interfaces. Proc Natl Acad Sci U S A 105: 530–535.
31. Tsai CJ, Lin SL, Wolfson HJ, Nussinov R (1997) Studies of protein-protein interfaces: a statistical analysis of the hydrophobic effect. Protein Sci 6: 53–64.
32. Ma B, Nussinov R (2007) Trp/Met/Phe hot spots in protein-protein interactions: potential targets in drug design. Current topics in medicinal chemistry 7: 999–1005.
33. Ma B, Elkayam T, Wolfson H, Nussinov R (2003) Protein–protein interactions: structurally conserved residues distinguish between binding sites and exposed protein surfaces. Proceedings of the National Academy of Sciences 100: 5772.
34. Richardson JS, Richardson DC (2002) Natural -sheet proteins use negative design to avoid edge-to-edge aggregation. Proceedings of the National Academy of Sciences 99: 2754.
35. Krishnan A, Zbilut JP, Tomita M, Giuliani A (2008) Proteins as networks: usefulness of graph theory in protein science. Curr Protein Pept Sci 9: 28–38.
36. Bode C, Kovacs IA, Szalay MS, Palotai R, Korcsmaros T, et al. (2007) Network analysis of protein dynamics. FEBS Lett 581: 2776–2782.
37. Martin AJ, Vidotto M, Boscariol F, Di Domenico T, Walsh I, et al. (2011) RING: networking interacting residues, evolutionary information and energetics in protein structures. Bioinformatics 27: 2003–2005.
38. Penel S, Hughes E, Doig AJ (1999) Side-chain structures in the first turn of the alpha-helix. J Mol Biol 287: 127–143.
39. Laemli U (1970) Cleavage of structural proteins during the assembly of the head of bacteriophage T4. Nature 227: 680–685.
40. Lesieur C, Cliff MJ, Carter R, James RF, Clarke AR, et al. (2002) A kinetic model of intermediate formation during assembly of cholera toxin B-subunit pentamers. J Biol Chem 277: 16697–16704.
41. Guex N, Peitsch MC (1997) SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. Electrophoresis 18: 2714–2723.
42. Murzin AG, Brenner SE, Hubbard T, Chothia C (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 247: 536–540.
43. Lo Conte L, Chothia C, Janin J (1999) The atomic structure of protein-protein recognition sites. J Mol Biol 285: 2177–2198.
44. Ma B, Nussinov R (2000) Molecular dynamics simulations of a [beta]-hairpin fragment of protein G: balance between side-chain and backbone forces1. Journal of molecular biology 296: 1091–1104.
45. Garratt RC, Thornton JM, Taylor WR (1991) An extension of secondary structure prediction towards the production of tertiary structure. FEBS letters 280: 141–146.
46. Minor DL, Jr., Kim P (1994) Measurement of the b-sheet-forming propensities of amino acids. Nature 367: 660–663.
47. FarzadFard F, Gharaei N, Pezeshk H, Marashi SA (2008) [beta]-Sheet capping: Signals that initiate and terminate [beta]-sheet formation. Journal of structural biology 161: 101–110.
48. Merkel JS, Sturtevant JM, Regan L (1999) Sidechain interactions in parallel [beta] sheets: the energetics of cross-strand pairings. Structure 7: 1333–1343.
49. Chakrabarti P, Janin J (2002) Dissecting protein-protein recognition sites. Proteins 47: 334–343.
50. Jones S, Thornton JM (1996) Principles of protein-protein interactions. Proc Natl Acad Sci U S A 93: 13–20.
51. Tacnet P, Cheong EC, Goeltz P, Ghebrehiwet B, Arlaud GJ, et al. (2008) Trimeric reassembly of the globular domain of human C1q. Biochim Biophys Acta 1784: 518–529.
52. Zrimi J, Ng Ling A, Giri-Rachman Arifin E, Feverati G, Lesieur C (2010) Cholera toxin B subunits assemble into pentamers - proposition of a fly-casting mechanism. PLoS One 5: e15347.
53. Dang LT, Purvis AR, Huang RH, Westfield LA, Sadler JE (2011) Phylogenetic and functional analysis of histidine residues essential for pH-dependent multimerization of von Willebrand factor. Journal of Biological Chemistry.
54. Lopez De La Paz M, Goldie K, Zurdo J, Lacroix E, Dobson CM, et al. (2002) De novo designed peptide-based amyloid fibrils. Proc Natl Acad Sci U S A 99: 16052–16057.
55. López De La Paz M, Serrano L (2004) Sequence determinants of amyloid fibril formation. Proceedings of the National Academy of Sciences of the United States of America 101: 87.
56. Marshall KE, Serpell LC (2009) Structural integrity of beta-sheet assembly. Biochem Soc Trans 37: 671–676.
57. Ruddock LW, Coen JJ, Cheesman C, Freedman RB, Hirst TR (1996b) Assembly of the B subunit pentamer of Escherichia coli heat-labile enterotoxin. Kinetics and molecular basis of rate-limiting steps in vitro. J Biol Chem 271: 19118–19123.
58. Conrady DG, Flagler MJ, Friedmann DR, Vander Wielen BD, Kovall RA, et al. (2010) Molecular basis of differential B-pentamer stability of Shiga toxins 1 and 2. PLoS One 5: e15153.
59. Moskaleva O, Melnik B, Gabdulkhakov A, Garber M, Nikonov S, et al. (2010) The structures of mutant forms of Hfq from Pseudomonas aeruginosa reveal the importance of the conserved His57 for the protein hexamer organization. Acta Crystallographica Section F: Structural Biology and Crystallization Communications 66: 760–764.
60. Nikulin A, Stolboushkina E, Perederina A, Vassilieva I, Blaesi U, et al. (2005) Structure of Pseudomonas aeruginosa Hfq protein. Acta Crystallographica Section D: Biological Crystallography 61: 141–146.
61. Panchenko AR, Luthey-Schulten Z, Wolynes PG (1996) Foldons, protein structural modules, and exons. Proc Natl Acad Sci U S A 93: 2008–2013.
62. Panchenko AR, Luthey-Schulten Z, Cole R, Wolynes PG (1997) The foldon universe: a survey of structural similarity and self-recognition of independently folding units. J Mol Biol 272: 95–105.
63. Mitraki A, van Raaij MJ (2005) Folding of beta-structured fibrous proteins and self-assembling peptides. Methods Mol Biol 300: 125–140.
64. Papanikolopoulou K, Teixeira S, Belrhali H, Forsyth VT, Mitraki A, et al. (2004a) Adenovirus fibre shaft sequences fold into the native triple beta-spiral fold when N-terminally fused to the bacteriophage T4 fibritin foldon trimerisation motif. J Mol Biol 342: 219–227.
65. Papanikolopoulou K, Forge V, Goeltz P, Mitraki A (2004b) Formation of highly stable chimeric trimers by fusion of an adenovirus fiber shaft fragment with the foldon domain of bacteriophage t4 fibritin. J Biol Chem 279: 8991–8998.
66. Zhang RG, Scott DL, Westbrook ML, Nance S, Spangler BD, et al. (1995) The three-dimensional crystal structure of cholera toxin. J Mol Biol 251: 563–573.
67. Shomura Y, Yoshida T, Iizuka R, Maruyama T, Yohda M, et al. (2004) Crystal structures of the group II chaperonin from Thermococcus strain KS-1: steric hindrance by the substituted amino acid, and inter-subunit rearrangement between two crystal forms. J Mol Biol 335: 1265–1278.
68. Gan L, Seyedsayamdost MR, Shuto S, Matsuda A, Petsko GA, et al. (2003) The immunosuppressive agent mizoribine monophosphate forms a transition state analogue complex with inosine monophosphate dehydrogenase. Biochemistry 42: 857–863.

Chapitre 8: Article publie: Intermolecular β-Strand Networks Avoid Hub Residues and Favor Low Interconnectedness: A Potential Protection Mechanism against Chain Dissociation upon Mutation

L'étude des réseaux d'interfaces β présentés dans des protéines oligomériques est poursuivie sur une base de données environ vingt fois plus grande que celle du chapitre 7, en termes de cas de protéines et contenant au total 1 048 interfaces β. Un des objectifs est de comprendre s'il existe des caractéristiques dans ces interfaces leur permettant de résister à des transitions conformationnelles entre différentes structures quaternaires.

En effet, certaines maladies sont liées à une transition conformationnelle de protéines oligomériques qui entraine une perte de leur fonction. En même temps, il existe beaucoup plus de protéines oligomériques contenant des interfaces β, qui ne change pas de conformations que de pathologies impliquant ces interfaces. Les propriétés des réseaux modélisant les 1 048 interfaces β de protéines non impliquées dans des pathologies sont donc étudies. Ses réseaux d'interfaces β suivent une distribution de degré de décroissance exponentielle qui semble pouvoir leur conférer une robustesse aux perturbations. La distribution exponentielle signifie que l'on n'a pas de distribution sans échelle avec des hubs et des bas degrés et qu'on n'est pas dans une situation ou la fragilité du réseau aux mutations viendrait de la mutation de hubs. Cependant on pourrait avoir une distribution exponentielle et avoir un réseau dont l'arrangement ne permettrait pas de résister à une perturbation (mutation ou conditions environnementales). Les réseaux β ont une connectivité faible car leurs acides aminés sont connectés qu'avec un petit nombre d'autres résidus du réseau. Cette caractéristique semble approprie pour protéger le réseau de la propagation de perturbation en son sein sous l'effet d'une mutation unique.

Pour tester cette hypothèse, le réseau de l'interface β du tétramère p53 a été étudié et comparé avec le prototype issu du set de données. La p53 est un trimère qui subit des changements conformationels sous l'effet de mutation, impliquées dans divers cancers. La comparaison montre une beaucoup plus grande connectivité du réseau de la p53 et des acides aminés de degrés plus élevés (donc des acides aminés avec plus de liens). La propensité d'un tel réseau à propager une perturbation a été testée avec la mutation G334V, mutation impliquée dans des cancers. Cette mutation unique entraine un réarrangement dans tout le réseau de l'interface de la p53, résultat appuyant l'hypothèse qu'une importante connectivité favorise la propagation de perturbations.

OPEN ACCESS Freely available online

PLOS ONE

Intermolecular β-Strand Networks Avoid Hub Residues and Favor Low Interconnectedness: A Potential Protection Mechanism against Chain Dissociation upon Mutation

Giovanni Feverati[1], Mounia Achoch[2], Laurent Vuillon[3], Claire Lesieur[4]*

1 Laboratoire d'Annecy-le Vieux de physique théorique (LAPTH UMR 5108), Université de Savoie, CNRS, Annecy le Vieux, France, **2** Laboratoire d'informatique systèmes, traitement de l'information et de la connaissance (LISTIC), Université de Savoie, Annecy le Vieux, France, **3** Laboratoire de mathématiques (LAMA UMR 5127), Université de Savoie, CNRS, Le Bourget du Lac, France, **4** Aging and imaging (AGIM FRE 3405), Université Joseph Fourier, CNRS, Grenoble, France

Abstract

Altogether few protein oligomers undergo a conformational transition to a state that impairs their function and leads to diseases. But when it happens, the consequences are not harmless and the so-called conformational diseases pose serious public health problems. Notorious examples are the Alzheimer's disease and some cancers associated with a conformational change of the amyloid precursor protein (APP) and of the p53 tumor suppressor, respectively. The transition is linked with the propensity of β-strands to aggregate into amyloid fibers. Nevertheless, a huge number of protein oligomers associate chains via β-strand interactions (intermolecular β-strand interface) without ever evolving into fibers. We analyzed the layout of 1048 intermolecular β-strand interfaces looking for features that could provide the β-strands resistance to conformational transitions. The interfaces were reconstructed as networks with the residues as the nodes and the interactions between residues as the links. The networks followed an exponential decay degree distribution, implying an absence of hubs and nodes with few links. Such layout provides robustness to changes. Few links per nodes do not restrict the choices of amino acids capable of making an interface and maintain high sequence plasticity. Few links reduce the "bonding" cost of making an interface. Finally, few links moderate the vulnerability to amino acid mutation because it entails limited communication between the nodes. This confines the effects of a mutation to few residues instead of propagating them to many residues via hubs. We propose that intermolecular β-strand interfaces are organized in networks that tolerate amino acid mutation to avoid chain dissociation, the first step towards fiber formation. This is tested by looking at the intermolecular β-strand network of the p53 tetramer.

Citation: Feverati G, Achoch M, Vuillon L, Lesieur C (2014) Intermolecular β-Strand Networks Avoid Hub Residues and Favor Low Interconnectedness: A Potential Protection Mechanism against Chain Dissociation upon Mutation. PLoS ONE 9(4): e94745. doi:10.1371/journal.pone.0094745

Editor: Frederique Lisacek, Swiss Institute of Bioinformatics, Switzerland

Received July 12, 2013; **Accepted** March 19, 2014; **Published** April 14, 2014

Funding: System Complex IXXI. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: claire.lesieur@agim.eu

Introduction

There exist proteins which function as oligomers by associating several copies of the same chains (homo-oligomers) or of different chains (hetero-oligomers). Chain association takes place through the formation of protein interfaces involving interactions between atoms of the amino acids of adjacent chains. Such intermolecular amino acid interactions are extensively studied by both experimental and computational approaches [1–5]. Alanine scanning mutagenesis have showed that only some of the amino acids of the interface account for the binding free energy [6]. Thus, there exists a subset of amino acids at interfaces, referred to as "hot spot" amino acids which are relevant for the chain association. This discovery has led to ample computational tool development aimed at identifying hot spots. The amino acids essential for interface formation are now known colloquially as hot spots, without necessarily implying alanine scanning validations.

Among proteins, some have the fold plasticity to undergo a transition from one oligomeric state to another. Of particular interest are the cases where the new oligomeric state impairs the protein function and leads to pathologies called protein misfolding diseases or conformational diseases. This transition is responsible for severe human diseases such as Alzheimer (Aβ-amyloid), Parkinson (synuclein) and cerebral amyloid angiopathy (cystatin C-amyloidosis). It is important to emphasize that the phenomenon is not restricted to neurodegenerative diseases but extends to cancer (p53), type II diabetes (IAPP, amylin), cardiovascular (transthyretin, serpin) and inflammatory diseases (Serpin) (reviewed in [7–11]). Note that in the previous sentence, for each of the diseases the protein undergoing the transition is indicated in brackets. A priori, these diseases are unrelated and the protein culprits do not share biological function, primary, secondary, tertiary or quaternary structures (initial or final). So the occurrence of the transition ought to be related to a local fold plasticity that allows transitions between different oligomeric states. It could be secondary structure plasticity as observed for the DIII loop of pore-forming toxins which becomes a β-hairpin and promotes the toxin's oligomerization or tertiary structure plasticity like the

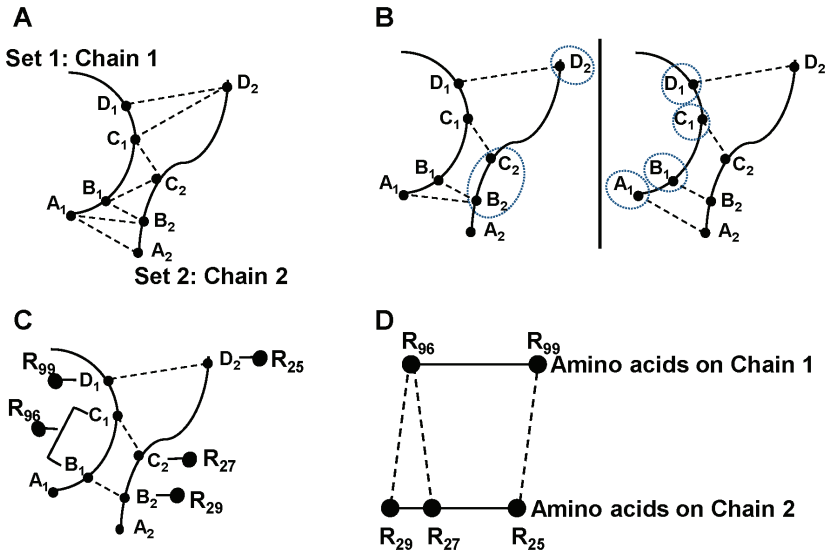


Figure 1. Illustration of the Gemini procedure on a trivial example. A. Interatomic distances between chain 1 and chain 2. On each chain, atoms are indicated by small filled circles labeled with letters. For clarity, only a few of the interatomic distances are indicated by dotted lines. B. Closest atoms. For every atom of S_1, Gemini chooses the closest atoms on S_2 (left picture) and for every atom of S_2, Gemini chooses the closest atoms on S_1 (right picture). The closest atoms are encircled. C. Mutually closest atoms. Gemini selects the atoms mutually the closest. The amino acids to which the mutually closest atoms belong are indicated by big filled circles. R stands for residue and the subscript is the position of the amino acid on the sequence. D. Gemini graph of amino acids in interaction. The distances between amino acids in contact are now arbitrary fixed to the same value because the information on the "real" interatomic distances is now lost. The pair of residues R99 and R25 is a single pair of amino acids ($k = 1$, that is one link connecting two residues). The residue R96 is a multiple contact amino acid because it is involved in two single pairs one with R29 and the other with R27, respectively.
doi:10.1371/journal.pone.0094745.g001

movement of the so-called "hinge loop" which leads to the formation of dimer or higher oligomeric states via a domain swapping mechanism [12–15].

The involvement of a local fold in the transition is in good agreement with the presence of a common structural motif in the pathological form of the culprit proteins. The pathological form, whether a fiber or an oligomer, involves interactions between two β-strands, each provided by a different chain (intermolecular β-strands). These intermolecular β-strands share several structural properties. They are recognized by the same antibody A11 [16]. Their formation depends on interactions between atoms of the backbone, result which has led to the proposal that aggregation is a generic property of the polypeptide chain [17,18]. They adopt a cross β structure which can be predicted from sequences by the PIRA (Parallel 'In Register' Arrangement) model, a network made of single pairs of residues [19–24]. Different predictors of the aggregation-prone sequences involved in the fiber formation are now available [25–30].

Nevertheless, intermolecular β-strands are common in protein oligomers that are not known to undergo a transition to pathological assemblies. This suggests that there is a protection mechanism that prevents some intermolecular β-strands from undergoing the transition. We are interested in identifying the features pertaining to the vulnerability of intermolecular β-strands to undergo a transition to pathological assemblies. The intermolecular β-strand interactions that occur in conformational diseases are often referred to as "aberrant" interactions because they lead to a loss of protein function and finally to the disease while the intermolecular β-strand interactions that occur in "healthy" protein oligomers are referred to as "functional" interactions.

Previous studies mainly in dimers have shown that the frequencies of individual amino acids in intermolecular β-strands and in intramolecular β-strands are not different [31]. Yet we have reported that intermolecular β-strands of oligomers of quaternary structures above dimer, have a scattered charge distribution in contrast to intramolecular β-strands and "aberrant" β-strands which have charges confined to their C- and N-terminal extremities [26,32,33]. Edge β-strands have charges centrally located which prevent their aggregation, explanation that holds for intermolecular β-strands as well [34]. In our study, the individual hot spots did not have any features that could account for a transition from "functional" to "aberrant" β-strand interactions. Because of the small size of the dataset (40 intermolecular β-strands), it was not possible to investigate the properties of the hot spot pairs or of the layout of the interactions between hot spots.

We have now built a larger dataset of 1048 intermolecular β-strands enabling us to explore such properties. The results show that the hot spots are not matched randomly but according to chemical and geometrical properties of the side chains of the amino acids. The role of the geometry is novel and might open new venues to apprehend how intermolecular β-strands are formed. The main result is that the interactions between hot spots are organized to resist to the effects of amino acid mutation, possibly avoiding in this way chain dissociation upon mutation, first step to fiber formation.

Results

The goal is to describe features of the hot spots involved in intermolecular β-strands and to consider how they may participate in a transition from "functional" to "aberrant" interactions. The intermolecular β-strands are represented as networks of hot spots in interaction with hot spots as nodes and interactions as links. Vocabulary related to graph and network theories are provided in methods.

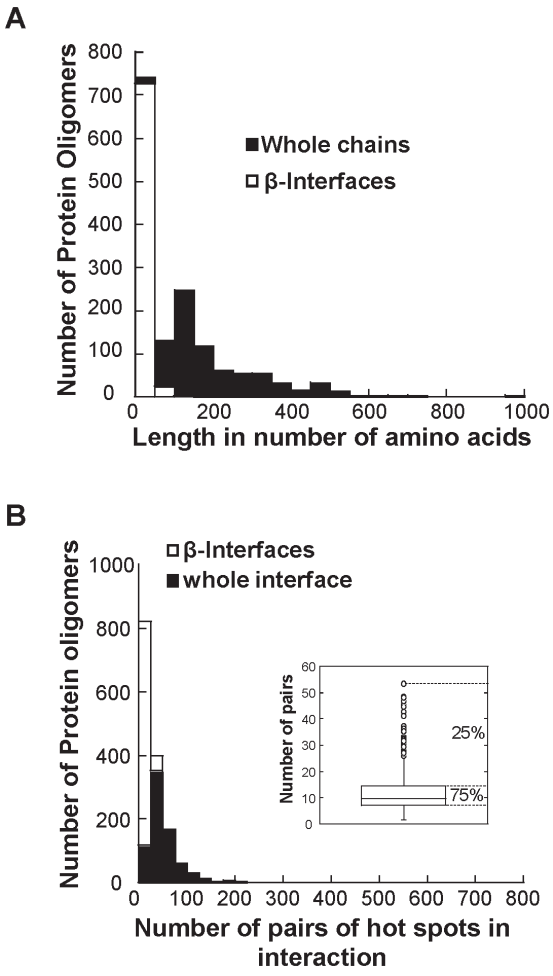


Figure 2. General features of the dataset. A. Histogram of the lengths (number of amino acids) of the whole chains (black bar) and of the intermolecular β-strands (white bar). B. Histogram of the number of hot spot pairs in the intermolecular β-strands (white bar) and in the whole interface (black bar). The inset is a box of the number of amino acid pairs in the intermolecular β-strands (quartile distribution). The values within the box (interquartile) represent 75% of the dataset. The points above the third quartile Q_3 (outside of the box) are β-interfaces whose number of amino acid pairs deviates significantly from the rest of the dataset.
doi:10.1371/journal.pone.0094745.g002

The tool Gemini

The nodes and the links of the networks are identified by our tool Gemini. Gemini has been described previously, hence we only briefly recall how the networks are built [35,36]. Each chain of a protein oligomer is considered as a set of points in the space whose positions are the Cartesian coordinates (x, y, z) of the atoms of the chain. The coordinates can be downloaded from the PDB. The atoms of the *chain 1* constitute the set 1 (S_1) and the atoms of the *chain 2*, the set 2 (S_2). Gemini calculates distances between every atom of S_1 and every atom of S_2 (interchain distances) but ignores the distances between atoms of a single set (intrachain distances) (Fig. 1A). Gemini chooses the closest atoms (Fig. 1B), and among them, retains only the pairs of mutually closest atoms (Fig. 1C). In other words, Gemini starts from an atom X_1 of S_1 and walks to its closest atom X_2 on S_2. It checks when coming back to S_1 by the shortest distance that it retraces its step to X_1. If not, the pair of atoms (X_1, X_2) is discarded, as for example for the pair (A_1, B_2) on figure 1C. The pairs of atoms that are mutually closest are considered to be interacting. At this stage the interchain interactions are symmetrical and the interface is referred to as around symmetrized [35]. In the last step, the pairs of atoms are replaced by their respective amino acids and a coarse-grained graph of amino acids in interaction is produced (Fig. 1D). Every amino acid has k interactions or k links where k equals to the number of atoms involved in a contact. There are single pairs of amino acids ($k = 1$, that is one link connecting two residues), multiple pairs of amino acids (k links connecting two residues) and multiple contact amino acids (an amino acid with k links to distinct amino acids).

Due to the choice of only mutually closest atoms, Gemini produces a graph of amino acids in interaction which is essentially a framework of interactions but not the set of all possible interactions. The amino acids selected by Gemini are detected as hot spots by available programs showing the robustness of defining an interface based only on geometry and its accuracy in picking up relevant amino acids [35]. It is important that Gemini does not need a cut-off distance to select atoms of the interface as classically done, for example to select preferentially backbone or side chain atoms. In this way Gemini avoids the variability of the selection inherent to the choice of a cut off [37]. Gemini naturally selects backbone and/or side chain atoms as part of the interface according to the geometry of the interface. Note that Gemini is applicable on any set of points in any metric space and can be used beyond the problem in question in the paper.

The dataset

The PDBs of 755 protein oligomers containing at least one intermolecular β-strand interface are extracted from the RCSB (Biological assembly) and in total 1048 intermolecular β-strand networks are constructed with Gemini. It is a non-redundant dataset of oligomers assembling three (trimer) to twelve subunits (dodecamers). The oligomers are selected only on the presence of intermolecular β-strands since we are looking for elements relevant to the formation of the interface but not to the formation of the whole chain. To fit that condition and alleviate the pressure of evolution due to fold or function similarities, we need a dataset with high diversities in terms of the features of the whole chains. The 755 protein oligomers classify into 234 SCOP families, 30 distinct functions, are produced by organisms from the three domains of life and have on average a full chain length of 206 ± 140 amino acids (average $\pm$ standard deviation) [38–40].

Now, on the contrary, we need a narrow diversity in terms of the features of the intermolecular β-strands to give evidences of a common construction mechanism. The average length of the intermolecular β-strands is 18 ± 13 amino acids, length calculated as the sum of the amino acids of the two β-strands. The distribution of the whole chain lengths is broader than that of the β-interface lengths (Fig. 2A). The intermolecular β-strands have on average 13 ± 8 hot spots, 75% have less than 16 hot spots and 25% have between 30 and 77 hot spots. Likewise, there are on average 12 ± 8 hot spot pairs per interface, 75% of the interfaces have less than 15 hot spot pairs while 25% have between 25 to 50 (Fig. 2B, inset). The number of hot spot pairs in the intermolecular β-strands is compared to the total number of hot spots pairs over the whole interfaces to assess the diversity of intermolecular β-strands in terms of the number of interactions necessary to build them. The distribution of the number of pairs in intermolecular β-strands is narrower than in the whole interface (Fig. 2B). Globally, 75% of the dataset have intermolecular β-strands sharing features. Moreover, there is no correlation between the length of the whole

Table 1. Whole chain amino acid and individual hot spot frequencies.

Amino acid	Whole chain	SC	BB
A	0.086	0.037	0.057
C	0.012	0.009	0.016
D	0.057	0.048	0.034
E	0.071	0.066	0.052
F	0.039	0.057	0.053
G	0.078	0.032	0.081
H	0.023	0.033	0.021
I	0.064	0.074	0.096
K	0.058	0.053	0.050
L	0.088	0.080	0.081
M	0.018	0.024	0.024
N	0.038	0.043	0.032
P	0.045	0.040	0.012
Q	0.033	0.040	0.037
R	0.051	0.062	0.050
S	0.056	0.062	0.061
T	0.056	0.075	0.070
V	0.083	0.090	0.109
W	0.011	0.019	0.013
Y	0.032	0.056	0.050

doi:10.1371/journal.pone.0094745.t001

chain and the length of the intermolecular β-strands (not shown, $R = 0.03$) supporting the idea that the two objects have independent features.

The dataset contains 568 anti-parallel β-sheets, 132 parallel β-sheets and 348 other β-strand arrangements (close packed β-strands) and 60% of the cases have β-strands with distinct sequences. One can already anticipate that the intermolecular β-strands of the dataset cannot be predicted based on a network of pairs of residues following a Parallel In Registered Arrangement (PIRA) because only 12% are parallel β-sheets and most β-strands have non identical sequences. The global features already highlight a network arrangement different from aggregation prone sequences [25].

Analysis of the properties of the residues in interaction in intermolecular β-strands

Gemini labels backbone and side chain atoms of the amino acids such that it produces two sub-graphs: one involving pure backbone interatomic interactions (BB networks) and the other involving interactions with at least one atom of the side chain (SC networks). We have shown that this distinction is necessary to exhibit features of intermolecular β-strands [32]. This is certainly related to the involvement of the backbone interactions in the hydrogen bond network of the β-sheets while in α-helices such backbone interactions are involved intramolecularly and are not interfering with intermolecular interactions. This is in good agreement with previous reports that side chain and backbone interactions are involved in hydrophobic and hydrogen bonding, respectively [1,41,42].

First, the properties of the individual hot spots are analyzed. Totals of 704623, 10692 and 5950 amino acids are observed in the whole chains, the SC and the BB hot spots, respectively. These figures give evidences of the reliability of the statistics which improves with the size of the sample. The amino acid frequencies are indicated in table 1 and used to measure the average chemical property, the global (GP) and local propensity (LP) of the amino acids (Tables 2 and 3, respectively). As observed previously the SC and BB hot spots have average chemical properties similar to the amino acids of the whole chains, global propensity and local amino acid distribution coherent with sequences made of β-strands as well as a scattered charge distribution [32]. Namely high β-sheet propensity residues (F, W, Y) are significantly more frequent while low β-sheet propensity (G and A) are significantly less [43–45]. The β-strand extremities are enriched in β-breaker amino acids (P and G) while high β-sheet propensity residues are enriched centrally (V, L) [46,47]. The charged residues R, K and E are enriched at the β-strand extremities whereas H and D residues are more frequent centrally when the local preferences of the SC charged residues is considered (Table 4).

Second, the properties of the pair of hot spots in interaction are analyzed. Because most of the intermolecular β-strands are not made of β-strands with an identical sequence, the occurrences n_{ab} and n_{ba} are initially counted but a χ^2 test calculated over the occurrences n_{ab} and n_{ba} shows that the differences are insignificant and so n_{ab} and n_{ba} occurrences are summed (Tables 5 and 6). The test ignored the values for the pair of identical residues for which a equals b. There are 10551 SC pairs and 5894 BB pairs, again highlighting the reliability of the statistics. The frequencies of the hot spot pairs f_{ab} are calculated with equation (1) and shown in the tables constituting the figure 3.

Table 2. Chemical properties of the intermolecular β-strands and of the whole chains (%).

Cases	Hydrophobic	Charged	Polar
Whole chain residues	49±5	26±5	25±6
BB hot spots (all)	58±27	19±20	23±23
BB hot spots (anti-parallel)	59±26	19±19	21±21
BB hot spots (parallel)	58±17	17±15	25±18
SC hot spots (all)	47±20	26±19	26±17
SC hot spots (anti-parallel)	47±21	26±19	26±18
SC hot spots (parallel)	46±17	25±16	29±14

doi:10.1371/journal.pone.0094745.t002

Table 3. Global propensities and local preferences of the SC and BB hot spots.

	SC hot spots				BB hot spots			
Amino acid	**Global propensity**	**Outer frequency (f_O)**	**Central frequency (f_C)**	**f_O-f_C**	**Global propensity**	**Outer frequency (f_O)**	**Central frequency (f_C)**	**f_O-f_C**
A	0.4	0.04	0.04	−0.003	0.7	0.06	0.05	0.009
C	0.8	0.01	0.01	−0.006	1.3	0.01	0.02	−0.012
D	0.8	0.05	0.05	−0.003	0.6	0.04	0.03	0.005
E	0.9	0.07	0.06	0.012	0.7	0.05	0.05	0.002
F	1.5	0.06	0.06	−0.004	1.4	0.05	0.06	0.003
G	0.4	0.03	0.03	0.002	1.0	0.09	0.07	0.02
H	1.4	0.03	0.03	−0.003	0.9	0.02	0.02	−0.004
I	1.1	0.07	0.08	−0.007	1.5	0.09	0.1	−0.004
K	0.9	0.06	0.05	0.014	0.9	0.06	0.05	0.011
L	0.9	0.08	0.08	0.001	0.9	0.07	0.09	−0.018
M	1.3	0.02	0.02	−0.003	1.3	0.02	0.03	−0.003
N	1.1	0.04	0.04	−0.002	0.8	0.04	0.03	0.013
P	0.9	0.05	0.03	0.017	0.3	0.02	0.01	0.009
Q	1.2	0.04	0.04	−0.002	1.1	0.04	0.04	0.005
R	1.2	0.08	0.05	0.025	1.0	0.05	0.05	0.006
S	1.1	0.06	0.07	−0.008	1.1	0.06	0.07	−0.012
T	1.4	0.07	0.08	−0.011	1.3	0.07	0.07	−0.004
V	1.1	0.08	0.1	−0.023	1.3	0.10	0.12	−0.019
W	1.7	0.02	0.02	−0.002	1.1	0.013	0.013	0.000
Y	1.8	0.06	0.05	0.005	1.6	0.05	0.05	−0,003
Average	1.1	0.05	0.05	0.000	1.0	0.05	0.05	0.000
S.D.	0.4	0.02	0.02	0.011	0.3	0.03	0.03	0.01

doi:10.1371/journal.pone.0094745.t003

Table 4. Local preferences of the charged amino acids in the SC hot spots.

Charged	Outer frequency (f_O)	Central frequency (f_C)	f_O-f_C
D	0.16	0.20	−0.042
E	0.25	0.25	0.004
H	0.11	0.14	−0.033
K	0.21	0.20	0.018
R	0.27	0.21	0.053
Average			0.000
S.D.			0.039

doi:10.1371/journal.pone.0094745.t004

$$(nab+nba)/\sum_{a=Y,b=Y}^{a=Y,b=Y}(nab+nba) \quad (1)$$

The ratio $(f_{ab}/f_a.f_b)$ is measured to compare observed values f_{ab} with expected values $(f_a.f_b)$ (Tables 7 and 8). If the frequency f_a is independent of the frequency f_b the ratio is equal to one. Overall the hot spots are not matched randomly since 70% and 66% of the BB and SC pairs, respectively, have a ratio that deviates from one. It is therefore necessary to measure the pair frequencies because they cannot be simply derived from the frequencies of individual hot spots.

To evaluate if the distinction between SC and BB hot spots is also relevant at the level of the pairs, the frequencies of the SC pairs are plotted against the frequencies of the BB pairs (Fig. 4). On the diagonal, there are 50 pairs out of a total of 210, thus indicating that 76% of the BB and SC pairs have different frequencies. It is therefore important to investigate them separately. Subsequent analyses are performed using quartiles to take into account the observation that 75% of the intermolecular β-strands share similar global interface features while 25% are more heterogeneous The amino acids with the highest 25% pair frequencies (> quartile Q_3) are considered as preferred contacts (Fig. 3, red) whereas those with the lowest 25% pair frequencies (< quartile Q_1) are considered as avoided contacts (Fig. 3, green). The neutral contacts have the frequencies between Q_1 and Q_3 (Fig. 3, white). The Q_3 and Q_1 of the SC hot spot pairs are 6.0×10^{-2} and 2.2×10^{-3}, respectively. The Q_3 and Q_1 of the BB hot spot pairs are 6.7×10^{-2} and 1.7×10^{-2}, respectively. In both networks, on average every amino acid pairs with 5 other types of amino acids out of its twenty pairing possibilities. The most preferred contacts are measured as amino acids which pair with a frequency above Q_3 with more than five other types of amino acids. For both networks, the most preferred contacts are I, L, V, S and T similarly to what was found for intermolecular β-strands in dimers [31]. On the other hand compared to the dimers F and Y residues are preferred in the SC networks while A and G are preferred in the BB networks, the residue E is preferred in both. Likewise, the most avoided contacts are measured as amino acids which pair with a frequency below Q_1 with more than five other types of amino acids. For both networks, the most avoided contacts are with C, M, W and H residues, similarly to intermolecular β-strands in dimers. In addition contacts with A, G and Q are avoided in the SC networks while contacts with N and Q residues are avoided in the BB networks.

The features of the hot spots pairs are then analyzed considering the chemistry and the geometry of amino acids (Tables 9 and 10, respectively). Both SC and BB hot spot pairs have similar tendencies for contacts with hydrophobic residues but the contacts with polar and charged residues are twice more frequent in the SC pairs. Even more blatant differences are the contacts between two charged residues, or between two polar residues or else between one charged and one polar residue, at least ten times more frequent in the SC networks. Considering geometrical properties (length of the side chains) the contacts with long and medium residues are significantly more frequent in the SC pairs than in the BB ones which on the contrary favor contacts between short side chain residues.

Third, the number of contacts of the hot spots is counted to determine whether the hot spots have multiple contacts. The BB networks have as many single contact hot spots (2941) as two contact hot spots (2993) but very little three contact hot spots (12). The degree distribution $P(k)$ is equal to the ratio of the number of hot spots with k contacts to the total number of hot spots. For the BB networks, $P(k)$ has a bell-like shape with an average $<k>$ contacts equals to 1.5 (Fig. 5A). On the other hand, $P(k)$ for the SC networks falls on a straight line when plotted on a linear-log scale indicating an exponential decay, a variation from the power law distribution observed for real networks [48] (Fig. 5A, $R^2=0.99$). The average $<k>$ contact of the SC hot spots is 1.4.

To determine a prototype intermolecular β-strand network, we use a binomial model with 9 amino acids per strand, 6 hot spots and the probability $p=0.16$ of having a contact (see methods for definition of a binomial law). These values are based on the averages of 18 amino acids, 12 hot spots and 10 links per interface measured over the dataset. A fully connected graph of 9 amino acids per strand (all amino acids have at least one link with all others) would have 81 links (9 by 9) and so in total on the dataset 84888 links. Only 13628 links are measured, thus the probability p of making a contact (having a link) is equal to 0.16 (13628/84888). Assuming that the amino acids have a uniform distribution of links (i.e. all amino acids have the same probability of making a link), the binomial model calculates a prototype network with 21% of non-connected amino acids (not hot spots), 36% of amino acids with one contact and 43% of amino acids with more than one contact, 27% of amino acids would have two contacts and 12% would have three. The observed data indicate 49% of amino acids with one contact and 51% amino acids with more than one contact, 33% with two, 14% with three and 4% with more than 3 contacts. The observed data are measured on hot spots only and so do not take into account the non-connected amino acids. In the binomial model, the "hot spots", namely the amino acids with a link are 79% (36% with one contact and 43% with more than

A

BB	A	C	F	G	I	L	M	P	V	W	D	E	H	K	R	N	Q	S	T	Y	
A	6.1E-03	3.7E-03	7.7E-03	8.3E-03	1.3E-02	1.2E-02	1.3E-03	2.0E-03	8.7E-03	1.8E-03	2.1E-03	2.8E-03	3.2E-03	5.5E-03	4.2E-03	3.4E-03	4.0E-03	5.2E-03	4.2E-03	1.1E-02	A
C		1.0E-03	2.9E-03	2.5E-03	4.2E-03	1.7E-03	0E+00	0E+00	5.1E-03	0E+00	3.4E-04	0E+00	5.9E-04	1.1E-03	1.7E-03	0E+00	3.4E-04	2.4E-03	1.1E-03	2.2E-03	C
F			6.7E-03	1.0E-02	8.0E-03	6.4E-03	3.5E-03	5.9E-04	8.8E-03	1.0E-03	2.5E-03	6.2E-03	3.1E-03	6.4E-03	4.6E-03	3.3E-03	1.5E-03	4.1E-03	9.9E-03	4.4E-03	F
G				9.3E-03	1.3E-02	1.1E-02	2.6E-03	6.3E-03	1.7E-02	1.7E-03	3.1E-03	6.8E-03	3.1E-03	5.0E-03	7.3E-03	4.4E-03	4.9E-03	1.1E-02	1.5E-02	7.0E-03	G
I					1.9E-02	1.9E-02	6.1E-03	1.0E-03	2.1E-02	2.5E-03	7.0E-03	9.0E-03	1.9E-03	8.7E-03	3.5E-03	2.5E-03	4.7E-03	1.3E-02	9.5E-03	9.1E-03	I
L						1.0E-02	4.8E-03	1.9E-03	2.4E-02	1.7E-03	3.3E-03	6.8E-03	3.0E-03	4.1E-03	4.8E-03	3.2E-03	5.8E-03	6.0E-03	9.6E-03	1.2E-02	L
M							2.7E-03	0E+00	5.7E-03	5.9E-04	1.0E-03	7.6E-04	0E+00	2.0E-03	3.0E-03	5.9E-04	1.1E-03	2.7E-03	3.8E-03	2.5E-03	M
P								3.4E-04	6.8E-04	5.9E-04	3.4E-04	3.0E-03	8.5E-04	6.8E-04	9.3E-04	6.8E-04	8.5E-04	6.8E-04	2.5E-04	1.4E-03	P
V									1.7E-02	3.1E-03	6.5E-03	9.5E-03	6.4E-03	8.7E-03	7.7E-03	8.1E-03	5.5E-03	9.9E-03	1.8E-02	1.3E-02	V
W										6.8E-04	3.4E-04	9.3E-04	5.1E-04	2.0E-03	1.4E-03	2.3E-03	5.9E-04	3.4E-04	9.3E-04	2.4E-03	W
D											2.3E-03	3.1E-03	3.4E-04	4.5E-03	7.0E-03	3.9E-03	2.3E-03	4.4E-03	5.2E-03	2.3E-03	D
E												2.4E-03	5.9E-04	1.1E-02	1.0E-02	3.8E-03	5.9E-03	8.8E-03	4.5E-03	4.2E-03	E
H													1.3E-03	2.0E-03	1.7E-03	6.8E-04	0E+00	3.2E-03	5.2E-03	2.1E-03	H
K														5.9E-03	3.2E-03	3.9E-03	2.3E-03	7.0E-03	3.6E-03	5.1E-03	K
R															5.1E-03	1.8E-03	7.8E-03	5.1E-03	1.0E-02	4.2E-03	R
N																1.9E-03	5.6E-03	2.8E-03	6.8E-03	1.6E-03	N
Q																	3.2E-03	7.6E-03	3.8E-03	2.0E-03	Q
S																		6.4E-03	7.8E-03	4.0E-03	S
T																			9.3E-03	6.2E-03	T
Y																				2.7E-03	Y

B

SC	A	C	F	G	I	L	M	P	V	W	D	E	H	K	R	N	Q	S	T	Y	
A	2.5E-03	1.2E-03	5.7E-03	1.0E-03	4.0E-03	5.5E-03	2.5E-03	2.5E-03	5.4E-03	2.6E-03	1.7E-03	2.1E-03	2.1E-03	2.3E-03	2.9E-03	3.4E-03	1.7E-03	3.0E-03	4.7E-03	7.1E-03	A
C		2.1E-03	8.1E-04	5.7E-04	1.8E-03	1.4E-04	2.7E-04	7.6E-04	2.3E-03	0E+00	3.8E-04	2.8E-04	2.2E-04	8.4E-04	6.6E-04	1.9E-04	2.4E-04	1.7E-03	3.0E-04	1.5E-03	C
F			1.1E-02	4.2E-03	9.8E-03	1.2E-02	3.8E-03	3.7E-03	1.1E-02	2.4E-03	1.4E-03	6.3E-03	1.7E-03	3.8E-03	4.1E-03	4.1E-03	3.1E-03	4.8E-03	8.2E-03	7.8E-03	F
G				0E+00	2.6E-03	3.7E-03	9.5E-04	3.3E-03	3.7E-03	1.4E-03	3.5E-03	1.1E-03	1.8E-03	3.9E-03	5.0E-03	3.2E-03	4.8E-03	3.9E-03	4.6E-03	3.0E-03	G
I					1.6E-02	1.6E-02	5.3E-03	3.8E-03	1.4E-02	1.6E-03	3.4E-03	5.5E-03	3.5E-03	5.3E-03	4.6E-03	3.9E-03	6.0E-03	6.6E-03	6.7E-03	7.8E-03	I
L						1.4E-02	3.6E-03	7.5E-03	1.8E-02	4.7E-03	4.2E-03	4.7E-03	4.2E-03	4.6E-03	5.9E-03	3.2E-03	5.1E-03	4.1E-03	1.1E-02	1.0E-02	L
M							3.5E-03	2.5E-03	6.0E-03	8.8E-04	2.0E-03	1.1E-03	6.2E-04	2.8E-03	2.3E-03	1.2E-03	1.2E-03	1.2E-03	1.6E-03	2.6E-03	M
P								6.0E-03	7.9E-03	1.9E-03	4.1E-03	2.7E-03	1.8E-03	2.4E-03	4.9E-03	1.6E-03	1.6E-03	4.6E-03	6.7E-03	4.3E-03	P
V									1.6E-02	3.4E-03	2.4E-03	7.0E-03	4.2E-03	5.7E-03	8.2E-03	5.4E-03	5.6E-03	8.3E-03	1.2E-02	1.0E-02	V
W										8.5E-04	4.5E-04	1.9E-03	1.3E-03	2.6E-03	1.9E-03	2.1E-03	5.6E-04	1.0E-03	4.6E-03	3.0E-03	W
D											3.3E-03	3.5E-03	4.3E-03	1.2E-02	1.8E-02	4.9E-03	3.5E-03	6.9E-03	8.6E-03	3.3E-03	D
E												5.2E-03	7.8E-03	1.9E-02	2.1E-02	6.0E-03	5.8E-03	1.0E-02	6.7E-03	6.1E-03	E
H													4.6E-03	2.7E-03	4.3E-03	3.2E-03	2.1E-03	3.9E-03	6.8E-03	4.0E-03	H
K														3.8E-03	4.2E-03	4.6E-03	4.3E-03	5.4E-03	6.5E-03	7.0E-03	K
R															5.3E-03	5.7E-03	4.8E-03	8.0E-03	8.2E-03	7.5E-03	R
N																5.5E-03	7.4E-03	6.4E-03	6.0E-03	3.2E-03	N
Q																	5.0E-03	5.2E-03	5.0E-03	3.8E-03	Q
S																		7.5E-03	1.1E-02	6.6E-03	S
T																			1.2E-02	8.2E-03	T
Y																				6.1E-03	Y

C

SC/BB	A	C	F	G	I	L	M	P	V	W	D	E	H	K	R	N	Q	S	T	Y	
A	0.4	0.3	0.7	0.1	0.3	0.5	1.9	1.3	0.6	1.5	0.8	0.8	0.7	0.4	0.7	1.0	0.4	0.6	1.1	0.7	A
C		2.0	0.3	0.2	0.4	0.1	n.a.	n.a.	0.5	n.a.	1.1	n.a.	0.4	0.8	0.4	n.a.	0.7	0.7	0.3	0.7	C
F			1.7	0.4	1.2	1.8	1.1	6.3	1.3	2.3	0.6	1.0	0.6	0.6	0.9	1.2	2.1	1.2	0.8	1.8	F
G				0.0	0.2	0.3	0.4	0.5	0.2	0.8	1.1	0.2	0.6	0.8	0.7	0.7	1.0	0.3	0.3	0.4	G
I					0.8	0.9	0.9	3.7	0.7	0.6	0.5	0.6	1.9	0.6	1.3	1.5	1.3	0.5	0.7	0.9	I
L						1.4	0.7	4.0	0.8	2.8	1.3	0.7	1.4	1.1	1.2	1.0	0.9	0.7	1.1	0.8	L
M							1.3	n.a.	1.1	1.5	2.0	1.5	n.a	1.4	0.8	2.1	1.1	0.4	0.4	1.0	M
P								17.6	11.6	3.2	12.1	0.9	2.1	3.6	5.3	2.4	1.9	6.8	26.4	3.2	P
V									1.0	1.1	0.4	0.7	0.7	0.6	1.1	0.7	1.0	0.8	0.7	0.8	V
W										1.2	1.3	2.0	2.6	1.3	1.3	0.9	0.9	3.1	4.9	1.2	W
D											1.5	1.1	12.6	2.6	2.6	1.3	1.5	1.6	1.7	1.4	D
E												2.2	13.1	1.7	2.1	1.6	1.0	1.1	1.5	1.4	E
H													3.6	1.4	2.5	4.7	n.a.	1.2	1.3	1.9	H
K														0.6	1.3	1.2	1.9	0.8	1.8	1.4	K
R															1.0	3.2	0.6	1.6	0.8	1.8	R
N																2.9	1.3	2.3	0.9	2.0	N
Q																	1.5	0.7	1.3	2.0	Q
S																		1.2	1.4	1.7	S
T																			1.3	1.3	T
Y																				2.2	Y

Figure 3. Tables of the f_{ab} pair frequencies. A. Observed BB pair frequencies. B. Observed SC pair frequencies. The frequency *fab* is for pairs of hot spots *ab* read on the lines *a* and the columns *b*. The preferred ($>Q_3$) and avoided ($<Q_1$) pairs are indicated by red and green color, respectively. The pairs with a frequency between Q_1 and Q_3 are not colored. The residues are ordered alphabetically within hydrophobic, charged and polar groups. C. SC and BB pair distinction. The ratios of the frequency of a pair *ab* in the SC sub-networks to its frequency in the BB sub-networks are indicated. The pairs more frequent in the SC sub-networks are indicated in red (ratio >1.2) and the pairs more frequent in the BB sub-networks are indicated in green (ratio <0.8). For ratio ranging from 0.8 to 1.2, the pairs are not colored. The abbreviation n.a. stands for "not applicable" which is division per zero, those pairs are more represented in the SC sub-networks.
doi:10.1371/journal.pone.0094745.g003

one). The percentage of amino acids with *k* contacts over a network made only of hot spots can be estimated for the binomial model by multiplying the calculated values by a factor 100/79. That produces 46% of hot spots with one contact (36 * 100/79), 54% (27*100/79) of hot spots with more than one contact, 34% (27 *100/79) with two, 15% (12 *100/79) with three and 5% of hot spots with more than three contacts in good agreement with observed values.

We then looked whether the hot spots had unusual amino acid features according to their number of contacts. The frequency of a hot spot in multiple contacts is divided by its frequency in single contact to measure the amino acid propensity to have multiple contacts. This propensity is plotted against the respective number of atoms of the side chain. No correlation is found for the BB hot spots (not shown, R = 0.41) and only branched residues V, I and L have a higher tendency of making two interactions suggesting that they are enriched in intermolecular β-strands involving parallel β-strands. On the other hand, there is a good linear correlation for the SC hot spots (Fig. 5B, R = 0.8). Thus, the propensity of the SC hot spot to make contacts is proportional to the number of its side chain atoms. Lastly, the probability of having hot spots with more than three contacts ($k>3$) is plotted against the number of atoms and compared to the probability of having a hot spot with one contact only (Fig. 5C). The probability of having hot spots with more than three contacts increases with the number of atoms whereas the probability of single hot spots distributes around a probability equal to 0.05. This probability (1/20) implies identical chance for all amino acids to have a single contact indicating no amino acid specificity for such contact number. On the other hand, only residues with more than 14 atoms (F, Y, R and W) have a probability above 0.05 to make more than three contacts, with the exception of the residue K.

Discussion

The analysis of the individual hot spot properties confirms a scattered charge distribution on the β-strands, high β-sheet propensity residues enriched centrally and more particularly branched side chain residues (V and L). This indicates that linear information, namely the information read on the sequence of the β-strands, codes essentially for solubility and regulation of the secondary structures.

Discriminating SC and BB interactions is again relevant at the level of the pairs as the SC and BB pair preferences diverge significantly. The ratio of SC and BB hot spots and the ratio of SC and BB pairs are on average around 2, indicating that the SC preferences are likely to have more influence over the intermolecular β-strands. One novel observation is that the pair matching is not only based on the chemistry of the amino acids but also on their geometry as seen in the preferences for long or charged residues in the SC pairs and for small or hydrophobic residues in the BB pairs. There is even enrichment in pairs combining amino acid properties such as pairs between long and charged residues or pairs between long and polar residues. In both SC and BB pairs, the branched residues V, I and L are preponderant contacts. A chemical-centric view for the pair matching is obviously ill-appropriate and in fact the pairing calls upon the versatile properties of amino acids. It might be interesting to explore the role of geometrical parameters on the formation of intermolecular β-strands, experimentally and theoretically. For instance, one theoretical approach would be to use Minimum Steiner trees which offer a purely geometrical description of the amino acids, to determine whether the pair matching yields a minimum energy conformation of the interface [49]. Contacts between identical residues represent only around 10% of the total preferred contacts indicating a minor role in the matching process. This differs from previous report on dimeric intermolecular β-strands and from the prediction by a PIRA model [25,31]. The data show that the 2D information, namely the amino acid pairing is not random and is important for the intermolecular β-strands, not surprisingly since β-strands are not viable without making interactions with another structural element.

Now the SC and BB networks do not differ only by their amino acid pairing but also by distinct network features. The BB networks have nodes with single or two contacts probably reflecting the hydrogen bond networks of anti-parallel (single contact) and parallel β-sheets (two contacts), respectively. The BB networks would essentially code for secondary structure interactions. The SC networks follow an exponential decay degree distribution and have nodes with one, two or three contacts but rarely with more than three. Thus the intermolecular β-strands result from the juxtaposition of two networks and the information for making the interface is encoded via a double layer of interactions. One layer is composed of the BB atoms and provides promiscuous interactions, namely low specificity in terms of amino acid composition and interaction motifs. The second layer is composed of the SC atoms which on the contrary provide selective interactions, high specificity in terms of amino acid composition and interaction motifs. Such type of double layer of interactions has been depicted for the interfaces between colicins and their immunity binding proteins as a way to evolve binding affinity [50]. There is also a precedent describing monomeric proteins and intramolecular amino acid interaction networks [51]. One network, based on short range interactions between Cα, had a bell curve degree distribution (random network feature) whilst the other based on long range interactions (side chain atoms) had an exponential decay degree distribution (single-scale network feature).

The exponential decay degree distribution likely fits a network optimized to reduce the number of links, relevantly because it costs to make a chemical link. Moreover, the data shows that above three contacts there is a strong stringency on the choice of the amino acids, suggesting that a node with too many links, a hub, would seriously decrease the sequence plasticity to successfully realize an interface. Intermolecular β-strands are very plastic in term of sequence requirement and seem therefore built to avoid hubs. Hubs are communication devices but also the Achilles' heel of the network: a modification of a hub spreads changes within the whole network because the hubs are connected to many nodes [52]. The propagation of changes upon node modification is called network rewiring [53]. The intermolecular β-strand networks which lack hubs are likely little inclined to rewiring because of

Table 5. Observed BB Pair occurrences.

BB	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	$\sum row$	$\sum$column	Average	Chi^2
A	36	11	5.5	7	23	24	8.5	39.5	16.0	34.0	4.0	10	5.5	11.5	13	15	12.5	25.5	5.5	32	303.0	309.7	306.3	0.8
C	11	6	1	0	8	7.5	2	12	2.5	5	0	0	0	1	5	6	3.5	16.5	0	6	8.7	89.5	88.3	0.9
D	7	1	13.5	10	8	8.5	1	23	15	10	3.5	11	1	6.5	20	13.5	15	21	1	8	184	168	176	0.4
E	9.5	0	8.5	14	19.5	20	1.5	27.5	31.5	19.5	2	12	8.5	17	30.5	26	14.5	28.5	3	13	292.5	285.8	289.2	0.8
F	22.5	9	7	17	39.5	28.5	8.5	23.5	19	19	9.5	10	1.5	4.5	14	11	29.5	26	3	13.5	276.5	285.7	281.1	0.7
G	24.7	7.5	10	20.3	32.7	54.7	9	37.5	14.5	32.8	8	13.5	18.5	15.2	21	34	48.8	51.8	5	22	426.8	405.8	416.3	0.5
H	10.5	1.5	1	2	9.5	9	7.5	5.5	6	8.5	0	2	2	0	5	10	14.5	19.5	2	6	114.5	111	112.8	0.8
I	39.5	13	18.5	25.5	23.5	41	5.5	111.5	23.5	52.5	18	6.5	3	14.5	8.5	37.5	28.5	60	7	27.5	453.5	474.5	464	0.5
K	16.5	4	11.5	34	19	15	5.5	27.5	35	12	6	12	2	7	9.5	18.5	11	27.5	6	15	259.5	252.5	256	0.8
L	35	5	9.5	20.5	19	32	9	57.5	12	61	13.5	9	5	16	13	18	28.5	70.5	5.5	36.5	415	412.7	413.8	0.9
M	3.5	0	2.5	2.5	11	7.5	0	18	6	15	16	2	0	3.5	9	7.5	11.5	16.5	2	8	126	122.5	124.3	0.8
N	10	0	12	10.5	9.5	12.5	2	8.5	11	10	1.5	11	2	16.5	5	8.5	20	26.5	7	5	178	172	175	0.7
P	6	0	1	9	2	18.5	3	3	2	6	0	2	2	2.5	3	2	1	2	2	4	69	64	66.5	0.7
Q	12	1	7	17.5	4.5	14	0	13	6.5	18	3	16.5	2.5	19	23.5	21	12	17	2	7	198	193.7	195.8	0.8
R	11.5	5	21	30	13	22	5	12	9.5	15	8.5	5.5	2.5	22.5	30	15	29	23.5	5	12.5	268	263	265.5	0.8
S	15.5	8	12.5	26	13	33.3	9	40.5	22.5	17.5	8.5	8	2	23.5	15	37.5	22.7	31.5	1	12.5	322.5	305.8	314.2	0.5
T	12.5	3	15.5	12	29	37	16	27.5	10	28	11	20	.5	10.5	30.5	23.5	55	54.5	3	18	362	374	368	0.7
V	26	13.5	17.5	27.5	26	51	18	64.5	24	70	17	21	2	15.5	22	27	50.5	99.5	8.5	39.5	541	566.7	553.8	0.4
W	5	0	1	2.5	3	5	1	8	6	4.5	1.5	6.5	1.5	1.5	3.5	1	2.5	9.5	4	7	70.5	75.5	73	0.7
Y	31.5	7	5.5	12	12.5	19.5	6.5	26	15	35.3	7	4.5	4	4.5	12	10.8	18.5	38.8	7	16	278	293	285.5	0.5

The values in a given row are the occurrences of the residue *a* in contact with the residues b_i, cited on columns. Thus, there are twenty rows a_i and twenty column b_i, - *i*- covering the 20 different amino acids. Due to the counting procedure the table is read row-wise (material and methods).

doi:10.1371/journal.pone.0094745.t005

Table 6. Observed SC Pair occurrences.

SC	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	$\sum row$	$\sum$column	Average	Chi^2
A	26.3	7.3	9.2	11.7	35	5	12.5	22.8	13	35	15	18	15	9.8	17	15.8	26.5	32.2	14.5	44	359.3	288.6	324	0
C	5.3	22	2	2	5	3	1.5	11	4.5	1	1.3	1	4.5	1	4	8.5	1.3	12	0	9	78	72.3	75.2	0.6
D	8.8	2	35.3	19.3	8.3	17	23	18.7	64.5	21.8	10.5	25.8	20.7	20	101.5	33.8	44	11.8	2.8	18.7	473.2	460.6	466.9	0.7
E	11	1	17.9	54.8	34.4	4.8	42.7	28.3	102.3	24	5.5	28.4	14.7	29.7	120	49.4	33.6	35.3	11.3	37	631.3	628.5	629.9	0.9
F	24.6	3.5	6.8	32.5	120	18.3	9.6	48.1	20.2	55.7	18.9	21.7	17.3	16.1	21.7	21.7	41.8	55.4	12.7	43.2	489.5	553.2	521.4	0
G	5.5	3	20	7	26.5	0	11	15.3	25	22.5	5.5	17.8	20	27.5	35.3	20.8	25.5	21	8.7	20	338	255.2	296.6	0
H	10	0.8	22.2	39.2	8.8	8.3	48.5	16.7	12.5	20.7	3	15.7	8.7	9.8	24.2	19.7	35.3	21.3	7.6	21.8	306.2	334.1	320.2	0.3
I	19.8	8.5	17.2	29.9	55.5	12.4	19.8	169.1	28	83.2	30.5	21.3	21.7	35.3	24.9	33	33.5	76.2	8.3	43.4	602.4	593	597.7	0.8
K	10.8	4.3	60	99.8	20.1	16.7	16.5	28.4	40.3	23.9	14.5	22.6	115	25	24	24.8	37.8	28.6	13.8	38.5	525.2	530	527.6	0.9
L	23	.5	23	25.4	68.3	16.1	24	89.5	25	149.9	18.8	17.5	41.2	26.3	35.7	18.3	56	94.6	29.7	57.3	690.1	662.4	676.2	0.5
M	11.1	1.5	11.1	6.4	20.9	4.5	3.5	25.2	14.9	18.8	36.8	6.1	12	7	11.5	5	8.9	27.3	4.3	15.2	215.2	231.7	223.5	0.4
N	17.7	1	25.7	34.7	21.7	16.1	18	20.2	25.5	16.1	6.8	57.6	8.5	38.7	32	32.8	31.5	26.7	12	15.8	401.4	398	399.7	0.9
P	11.2	3.5	22.7	13.5	22.2	14.3	10	18.5	10.8	38.3	14.3	8.8	63	8.5	28	22	33.5	41.7	11.8	25.8	359.5	364.6	362	0.8
Q	8.5	1.5	17.2	31.8	17	22.7	12	28.2	20.2	27.1	5.7	39.3	8.2	52.3	26.5	25.3	24.5	25	3.3	20.2	364.2	392.6	378.4	0.3
R	13.3	3	88.7	106.4	21.8	17.5	21.3	23.4	20.3	26.7	12.5	28	24.2	24.1	56.3	38.3	40.3	39.3	10	39.7	598.7	694.8	646.7	0
S	15.7	9.5	39	57.3	28.5	20.5	22	36.8	32.5	24.5	7.8	34.3	27	29.5	46	79.5	58.5	45.3	6.5	39.3	580.5	500.3	540.4	0
T	22.7	1.8	46.5	36.7	44.8	22.6	36.7	37.5	30.8	56.9	7.9	32.2	37.3	27.8	46.5	54.3	127.1	65.5	22.8	46	677.4	661.2	669.3	0.7
V	25.3	12.5	13.8	38.8	62.7	17.8	23.2	76.6	31.2	97.9	36	30.8	41.3	34	47.3	42.4	63.2	170.8	21.3	56.3	772.3	723.5	747.9	0.2
W	13.4	0	2	8.7	12.5	6.3	6.2	8.6	13.2	20.1	5	10.5	8.1	2.5	9.7	4.5	25.4	14.6	9	15.3	186.5	217.5	202	0.1
Y	30.8	7	15.8	27.2	39.2	11.5	20.8	39.4	35.6	48.2	12.1	18.3	19.3	20	39.1	29.8	40.1	49.8	16	64.1	519.9	606.4	563.2	0

The values in a given row are the occurrences of the residue a in contact with the residues b_i, cited on columns. Thus, there are twenty rows a_i and twenty column b_i, - i- covering the 20 different amino acids. Due to the counting procedure the table is read row-wise (material and methods).

doi:10.1371/journal.pone.0094745.t006

Table 7. Ratio of $f_{ab}/(f_a.f_b)$ for the BB hot spot pairs.

$f_{ab}/f_a.f_b$	A	C	F	G	I	L	M	P	V	W	D	E	H	K	R	N	Q	S	T	Y
A	2	4	3	2	2	3	1	3	1	2	1	1	3	2	1	2	2	1	1	4
C		4	3	2	3	1	0	0	3	0	1	0	2	1	2	0	1	3	1	3
F			2	2	2	1	3	1	2	2	1	2	3	2	2	2	1	1	3	2
G				1	2	2	1	6	2	2	1	2	2	1	2	2	2	2	3	2
I					2	2	3	1	2	2	2	2	1	2	1	1	1	2	1	2
L						2	2	2	3	2	1	2	2	1	1	1	2	1	2	3
M							5	0	2	2	1	1	0	2	2	1	1	2	2	2
P								2	1	4	1	5	3	1	2	2	2	1	0	2
V									1	2	2	2	3	2	1	2	1	2	2	2
W										4	1	1	2	3	2	6	1	0	1	4
D											2	2	0	3	4	4	2	2	2	1
E												1	1	4	4	2	3	3	1	2
H													3	2	2	1	0	3	4	2
K														2	1	2	1	2	1	2
R															2	1	4	2	3	2
N																2	5	1	3	1
Q																	2	3	1	1
S																		2	2	1
T																			2	2
Y																				1

doi:10.1371/journal.pone.0094745.t007

Table 8. Ratio of $f_{ab}/(f_a.f_b)$ for the SC hot spot pairs.

$f_{ab}/f_a.f_b$	A	C	F	G	I	L	M	P	V	W	D	E	H	K	R	N	Q	S	T	Y
A	2	3	3	1	1	2	3	2	2	4	1	1	2	1	1	2	1	1	2	3
C		24	1	2	3	0	1	2	3	0	1	0	1	2	1	0	1	3	0	3
F			3	2	2	3	3	2	2	2	1	2	1	1	1	2	1	1	2	2
G				0	1	1	1	3	1	2	2	1	2	2	3	2	4	2	2	2
I					3	3	3	1	2	1	1	1	1	1	1	1	2	1	1	2
L						2	2	2	3	3	1	1	2	1	1	1	2	1	2	2
M							6	3	3	2	2	1	1	2	2	1	1	1	1	2
P								4	2	3	2	1	1	1	2	1	1	2	2	2
V									2	2	1	1	1	1	1	1	2	1	2	2
W										2	1	2	2	3	2	3	1	1	3	3
D											1	1	3	5	6	2	2	2	2	1
E												1	4	5	5	2	2	2	1	2
H													4	2	2	2	2	2	3	2
K														1	1	2	2	2	2	2
R															1	2	2	2	2	2
N																3	4	2	2	1
Q																	3	2	2	2
S																		2	2	2
T																			2	2
Y																				2

doi:10.1371/journal.pone.0094745.t008

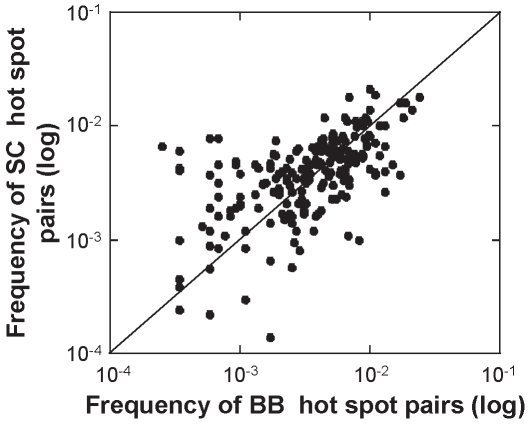


Figure 4. Comparison of the 210 frequencies of the BB and SC hot spot pairs. The frequencies of the SC hot spots pairs are plotted against those of the BB hot spots pairs, both in log scale. Pairs with identical BB and SC frequencies are on the diagonal. Pairs more frequent in SC are found above the diagonal whereas pairs more frequent in BB are found below the diagonal.
doi:10.1371/journal.pone.0094745.g004

their low interconnectedness. Counterintuitively, the robustness of intermolecular β-strands would appear based on a weak occurrence of links maintaining high sequence plasticity, cutting costs in term of links and reducing their vulnerability to changes (mutation).

It is tempting to speculate that a higher number of links is one of the necessary conditions to have a transition from "functional" to "aberrant" intermolecular β-strands. It is possible that "healthy" protein oligomers which become pathological fibers have interfaces with more links per nodes and networks more sensitive to rewiring than those which do not form fibers. To examine such possibility, the tumor suppressor p53 tetramer (PDB 1SAK, fig. 6A), a known case of healthy oligomer undergoing a transition to a fiber is considered. First, the Gemini graph of the WT p53 is generated (Fig. 6B). The greater occurrence of multiple contact residues is striking in the WT p53 network, supporting the hypothesis. The p53 hot spots have on average $<k> = 3$ contacts, twice the $<k>$ value of the intermolecular β-strand networks. The p53 network has 33% hot spots with more than three contacts which is 6 times more than the prototype network. On the other hand, it has 25% of single contact hot spots twice less than the prototype network. Consequently the interconnectedness is larger in the p53 network than in the prototype network.

To look at the sensitivity of the p53 network to single point mutation, the G334V mutant, a familial mutation that leads to the dissociation of the p53 tetramer, misfunctions of the protein and cancer development, is considered [54]. The Gemini graph of G334V is generated and network rewiring is investigated (Fig. 6C). The mutation has a strong global effect on the network as all the residues of the p53 intermolecular β-strands from 324 to 334 have their links modified by the mutation even when they are not directly linked to the residue 334. The modifications are either: (i) vanishing of the links (e.g. D324, G325), (ii) changes of the type of links such as side chain to backbone (e.g. I332, L330), (iii) decrease of the number of contacts (e.g. Q331, T329) or else (iv) changes of contacts (R333). The changes in the network are not limited to residues of the intermolecular β-strands but extend to interactions between residues that belong to α-helices. This definitely shows that there is significant network rewiring in p53 due to a single node modification, the mutation of the residue G334, again supporting the hypothesis. Mutation of other p53 residues such as T329A or Q331A also leads to similar network rewiring (not shown) which therefore cannot explain the capacity of the mutant G334V to form a fiber, because the T329A and Q331A mutants do not make to fiber [54]. The extent of the changes in the network might be such that the intermolecular β-strand interactions are destabilized promoting chain dissociation, the first step to fiber formation.

Conclusion. The key results are: (i) little information is accessible from individual amino acids (i.e. in sequences) and it is the pairs of amino acids that need to be investigated, (ii) the geometry of the amino acid side chains, so far neglected, is a key parameter to understand pair matching and finally (iii) intermolecular β-strands need to be further explored in terms of networks. The intermolecular β-strand networks are rather disconnected networks with no hubs but nodes with few links instead. Such a layout has several advantages as already discussed but probably the most relevant one is the secluding characteristic of the network which may well serve to limit the spread of changes, namely the rewiring, and protect the interface from dissociation upon mutation.

Table 9. SC and BB hot spot pair chemical tendencies.

Pair property	Total	SC tendency	BB tendency	Neutral
(Fhi, X)	155	58	65	32
(Ch, X)	90	50	26	14
(P, X)	90	46	22	22
(Fhi, Fhi)	55	24	23	8
(Fhi, Ch)	50	19	23	8
(Fhi, P)	50	15	19	16
(Ch, Ch)	15	12	1	2
(Ch, P)	25	19	2	4
(P, P)	15	12	1	2

The number of pairs with a ratio SC pair frequency to BB pair frequency above 1.0 ± 0.2 indicates the SC pair tendency. The number of pairs with a ratio below 0.8 ± 0.2 indicates the BB pair tendency (table based on Fig. 2C). The second column, total, indicates the pair combinatory of the chemical pair property mentioned in the first column. Fhi, Ch and P stand for hydrophobic, charged and polar residues. X stands for fhi, ch and P.
doi:10.1371/journal.pone.0094745.t009

Table 10. SC and BB hot spot geometrical pair tendencies.

Geometrical pair property	Total	SC preferred	BB preferred	Neutral
(L, X)	74	43	10	21
(M, X)	144	72	41	31
(S, X)	119	41	49	29
(L, L)	10	8	1	1
(L, M)	36	23	5	8
(L, S)	28	12	4	12
(M, M)	45	29	6	10
(M, S)	63	20	30	13
(S, S)	28	9	15	4

Legend as in table 9. L, M and S stand for long, medium and short side chains. × stands for L, M and S.
doi:10.1371/journal.pone.0094745.t010

Methods

Definitions

Graph. graph, or a network, is a set of many components that interact with each other through pairwise interactions. At a highly abstract level, the components can be reduced to a series of nodes that are connected to each other by links, with each link representing the interactions between two components. The nodes and links together form a network, or, in more formal mathematical language, a graph [55]. The terms nodes and links used in graph theory are amino acids/hot spots and contacts/interactions, respectively, in the present context. The number of links of a node is the degree *k* of the node. In the networks of hot spots in interaction, the residues are connected through different motifs. Two residues connected by only one link make a single pair while two residues connected by more than one link make a multiple pair. Hot spots involved in single pair are single contact hot spots. Hot spots with more than one individual contact are called multiple contact hot spots.

Global propensity (GP). The global propensity of an amino acid is the ratio of its frequency in a defined environment by its frequency in a database. Here the global propensity measures the frequency of every amino acid in intermolecular β-strands divided by its frequency in the whole chain.

Local preferences: the local amino acid preferences measure the preferred position of every amino acid on the β-strands. It is calculated as the difference of the frequency of a hot spot at the β-strand extremities (outer position) and its frequency when centrally located (any other position) on the strand.

Chemistry of the side chain of amino acid: charged amino acids are D, E, H, R and K; polar amino acids are N, Q, S, T and Y; hydrophobic residues are A, C, F, G, I, L, M, P, V and W.

Length of the side chain of amino acid: long side chain residues are K, W, R and K; medium side chain residues are D, N, L, I, H, E, Q, M and F and short side chain amino acids are G, A, P, C, S, T and V.

Methods

Construction of a non-redundant dataset

The Protein Data Bank (PDB) was first screened at the Research Collaboratory for Structural Bioinformatics (RCSB) for protein oligomers of stoichiometry above 2 and lower or equal to 12 [56]. Above dodecamers the number of cases becomes small for statistical analysis. Dimers are excluded from the dataset because of their diversity of orientation contacts implying broad diversity in recognition contact modes [57]. Viral and membrane proteins have been removed because they are likely to follow a different mechanism of interface formation than soluble oligomers. The coordinates of biological assembly were taken to select for non-crystallographic oligomers. NMR and X-ray structures are taken into account. PDB entries containing only backbone (BB) atoms, or only a few side-chain (SC) atoms, are discarded by monitoring the ratio of available SC and BB atoms for each of the twenty amino acids. Proteins with sequences similar at 90% identity are removed. As a result, 6234 PDBs have been tentatively treated with Gemini to describe the whole interface. There is a small minority of cases where Gemini stops before yielding the interface. Mainly, this is due to the presence of a single subunit in the PDB file, while Gemini expects several. This happens even if biological assemblies were downloaded from the RCSB. At this point, the interface is available for a set of 5248 proteins. Receptor-ligand, enzyme-inhibitor, and antigen-antibody types of interactions involve different ranges of K_D than permanent oligomers and as such are expected to have different recognition modes [42]. Therefore they are discarded from the dataset by removing proteins having at least one very short chain ($\leq$20 amino acids). Truncated proteins were also discarded from the dataset by selecting only cases having chains less than 20 amino acid different in length.

Using the secondary structure annotation provided in the PDB file, the cases with intermolecular β-strands were extracted according to the following set of rules (to be simultaneously satisfied): 1) at least 3 bonds must be between amino acids belonging to β-strands; 2) at least 2 interface amino acids of each subunit must be in a β-β bond; 3) at least 5 interface amino acids must be classified β. The first rule is actually redundant as it is implied by the second and the third. To simplify the treatment, in the case of hetero-oligomers with more than one intermolecular β-strand, only one, randomly chosen, has been considered. The final list has been screened against redundancies by mapping each PDB code into a UniProt identifier. This allows using the appropriate UniProt algorithms to find and remove redundant cases. After this final suppression, we are left with 755 proteins having 1048 regions of intermolecular β-strands.

Hot spots in interaction

A pair of hot spots is made of a hot spot –*a*- interacting with a hot spot –*b*-. Some hot spots participate in more than one pair at the same time and it is necessary to avoid their multiple counting.

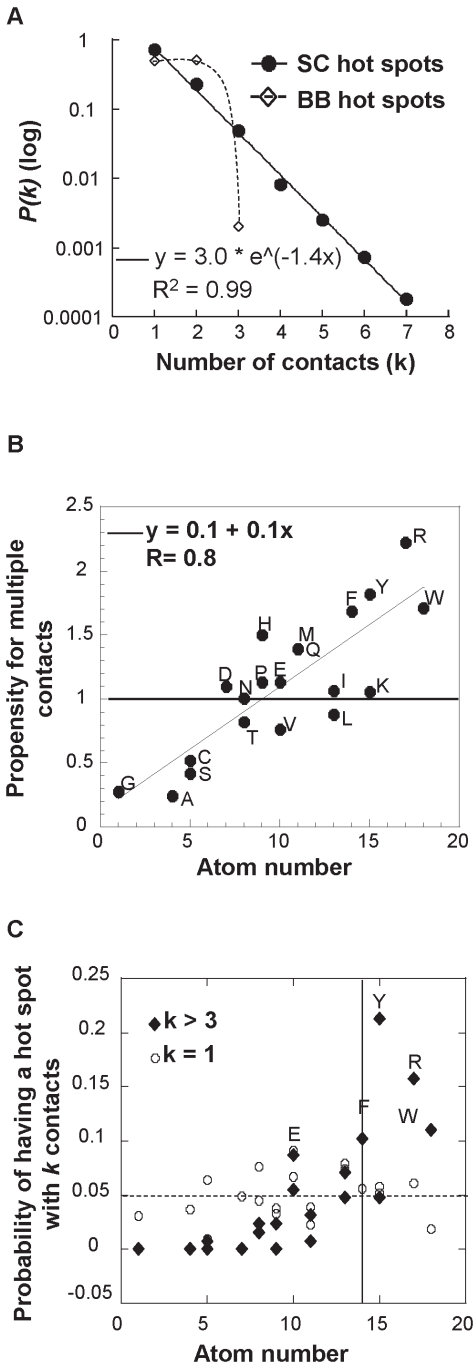


Figure 5. Number of contacts of the hot spots. A. The degree distributions of the BB and SC hot spots are plotted on a semi-log scale. The degree distribution *P(k)* of the SC hot spots decreases exponentially ($R^2 = 0.99$). B. Linear correlation between the number of atoms of a SC hot spot and its tendency to have more than one contact. The ratio of the frequency of an amino acid in multiple contacts to its frequency in single contact is plotted against the number of its side chain atoms. C. Probability of a SC hot spot to have *k* contacts. The probabilities for a SC hot spot to have $k>3$ (♦) or $k=1$ (○) are plotted against the number of atoms of its respective amino acid. The horizontal line indicates the probability at which every amino acid has the same probability to have *k* contacts (0.05 = 1/20). The vertical line indicates a number of atoms equals to 14.
doi:10.1371/journal.pone.0094745.g005

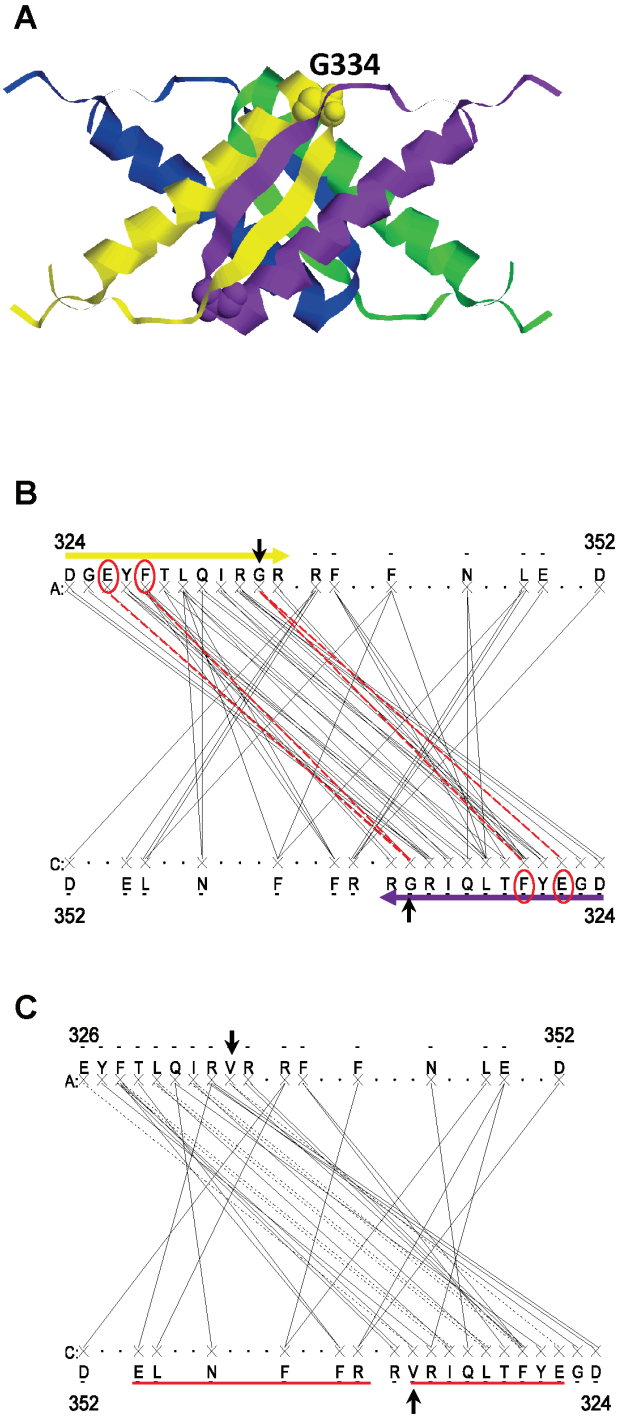


Figure 6. The p53 intermolecular β-strand network. A. Atomic structure of the p53 tetramerization domain (PDB 1SAK). The picture is generated with Rasmol, the four chains are shown in different colored ribbons. The G334 residue is indicated in spacefill. B. Gemini graph of the WT p53 tetramerization domain. The intermolecular β-strands composed of the residues 324 to 334, are highlighted by the yellow and purple arrows. The vertical arrows point to the residue 334. The links and hot spot contacts of G334 are shown by dotted red lines and red circles, respectively. C. Gemini graph of the G334V mutant. The hot spots whose links are affected by the mutation are underlined in red. The changes are not limited to residues in direct contact with G334 or to residues of the intermolecular β-strands.
doi:10.1371/journal.pone.0094745.g006

A pair (A1, A2) is counted 1/n time with *n* the number of bonds of A1. Let's consider a hot spot G forming a pair with T and another pair with L. Each of the (G, T) and (G, L) pairs is counted a half so the occurrence of G is equal to one and not to two if the pairs (G, T) and (G, L) had been counted one each instead of a half. This counting procedure implies that the tables of occurrences must be read row-wise (Tables 5 and 6). Now, when the number of interactions (bonds) issued from a hot spot is counted instead of the pair occurrences such normalization is unnecessary.

Statistical tools

χ^2. n_{ab} and n_{ba} pair occurrences. The total observed pair occurrences n_{ab} and n_{ba} are calculated for each residue as the sum of the occurrences on a row and the sum of the occurrences on a column (Tables 5 and 6 for the BB and SC sub-networks, respectively). The significance of the differences of the occurrences n_{ab} and n_{ba} was assessed using a χ^2 (equation 2) with one degree of freedom calculated as follows:

$$\chi^2 = \sum_{\substack{i=A,Y \\ j-A,Y}} \sum_{\substack{i=A,Y \\ j=A,Y}} (Oij - Eij)^2 / Eij \qquad (2)$$

With O_{ij} the observed occurrences (line i and column j on the tables 5 and 6) and E_{ij} the expected occurrences calculated as the average value of the total observed pair occurrences n_{ab} and n_{ba}. The sums are for the n_{ab} and the n_{ba} occurrence values. For one degree of freedom, a χ^2 value inferior to 3.84 is not significant (5% threshold significance).

Observed (f_{ab}) and expected values ($f_a \times f_b$). The significance of the differences of the observed (f_{ab}) and expected pair frequencies ($f_a \times f_b$) was also assessed using a χ^2 with O_{ij} and E_{ij} the observed and expected pair frequencies, respectively. This time it is calculated over a matrix where low occurrences (below 5) are summed and a p-value is calculated.

Binomial law. This law calculates the probability of making a link $P(k)$ over a large number of test n with p the probability to make a link and (1-p) the probability to make no link (equation 3). Thus the probability of any SC hot spot to make k links (i.e. k number of contacts) is calculated as the product of the probability for any node to make k links by its probability to make no link over n trials. When the calculated values are close to the observed values, the binomial law is a good model for estimating the number of links of the hot spots.

$$P(x=k) = \binom{n}{k} p^k (1-p)^{n-k} \qquad (3)$$

Virtual mutation

Fold X is used to generate the virtual mutation G334V in the PDB of the p53 tetramerization domain was designed following instruction in [58,59].

Availability of supporting data

The list of the 755 PDB cases and their respective intermolecular β-strands are available on request.

Acknowledgments

We thank the federation of research MSIF (Modelisation, Simulation, Interactions Fondamentales) for supporting our work (?http://lapth.cnrs.fr/msif/fr/). Mounia Achoch is founded by the region Rhone-Alpes. We thank Alain Henaut and Alexander Grossmann for critical reading of the manuscript. We thank Kave Salamatian for stimulating and useful discussions on networks and graph theory.

Author Contributions

Conceived and designed the experiments: GF LV CL. Performed the experiments: MA GF. Analyzed the data: CL LV. Contributed reagents/materials/analysis tools: GF LV. Wrote the paper: CL.

References

1. Janin J, Bahadur RP, Chakrabarti P (2008) Protein-protein interaction and quaternary structure. Q Rev Biophys 41: 133–180.
2. Tuncbag N, Kar G, Keskin O, Gursoy A, Nussinov R (2009) A survey of available tools and web servers for analysis of protein–protein interactions and interfaces. Briefings in Bioinformatics 10: 217.
3. DeLano WL (2002) Unraveling hot spots in binding interfaces: progress and challenges. Current opinion in structural biology 12: 14–20.
4. Talavera D, Robertson DL, Lovell SC (2011) Characterization of protein-protein interaction interfaces from a single species. PloS one 6: e21053.
5. Bahadur RP, Chakrabarti P, Rodier F, Janin J (2004) A dissection of specific and non-specific protein-protein interfaces. J Mol Biol 336: 943–955.
6. Clackson T, Wells JA (1995) A hot spot of binding energy in a hormone-receptor interface. Science 267: 383–386.
7. Eisenberg D, Jucker M (2012) The Amyloid State of Proteins in Human Diseases. Cell 148: 1188–1203.
8. Lomas DA, Carrell RW (2002) Serpinopathies and the conformational dementias. Nature Reviews Genetics 3: 759–768.
9. Cheng P-N, Pham JD, Nowick JS (2013) The Supramolecular Chemistry of β-Sheets. Journal of the American Chemical Society.
10. Bellotti V, Chiti F (2008) Amyloidogenesis in its biological environment: challenging a fundamental issue in protein misfolding diseases. Current opinion in structural biology 18: 771–779.
11. Ochieng J, Chaudhuri G (2010) Cystatin superfamily. J Health Care Poor Underserved 21: 51–70.
12. Iacovache I, Paumard P, Scheib H, Lesieur C, Sakai N, et al. (2006) A rivet model for channel formation by aerolysin-like pore-forming toxins. The EMBO Journal 25: 457–466.
13. Picone D, Di Fiore A, Ercole C, Franzese M, Sica F, et al. (2005) The role of the hinge loop in domain swapping. The special case of bovine seminal ribonuclease. J Biol Chem 280: 13771–13778.
14. Bennett MJ, Schlunegger MP, Eisenberg D (1995) 3D domain swapping: a mechanism for oligomer assembly. Protein Sci 4: 2455–2468.
15. Benito A, Laurents DV, Ribo M, Vilanova M (2008) The structural determinants that lead to the formation of particular oligomeric structures in the pancreatic-type ribonuclease family. Curr Protein Pept Sci 9: 370–393.
16. Kayed R, Head E, Thompson JL, McIntire TM, Milton SC, et al. (2003) Common structure of soluble amyloid oligomers implies common mechanism of pathogenesis. Science 300: 486–489.
17. Guijarro JI, Sunde M, Jones JA, Campbell ID, Dobson CM (1998) Amyloid fibril formation by an SH3 domain. Proc Natl Acad Sci U S A 95: 4224–4228.
18. Dobson CM (1999) Protein misfolding, evolution and disease. Trends Biochem Sci 24: 329–332.
19. Petkova AT, Ishii Y, Balbach JJ, Antzutkin ON, Leapman RD, et al. (2002) A structural model for Alzheimer's beta -amyloid fibrils based on experimental constraints from solid state NMR. Proc Natl Acad Sci U S A 99: 16742–16747.
20. Der-Sarkissian A, Jao CC, Chen J, Langen R (2003) Structural organization of alpha-synuclein fibrils studied by site-directed spin labeling. J Biol Chem 278: 37530–37535.
21. Kajava AV, Aebi U, Steven AC (2005) The parallel superpleated beta-structure as a model for amyloid fibrils of human amylin. J Mol Biol 348: 247–252.
22. Krishnan R, Lindquist SL (2005) Structural insights into a yeast prion illuminate nucleation and strain diversity. Nature 435: 765–772.
23. Margittai M, Langen R (2004) Template-assisted filament growth by parallel stacking of tau. Proc Natl Acad Sci U S A 101: 10278–10283.
24. Lv G, Kumar A, Giller K, Orcellet ML, Riedel D, et al. (2012) Structural comparison of mouse and human alpha-synuclein amyloid fibrils by solid-state NMR. J Mol Biol 420: 99–111.
25. Trovato A, Chiti F, Maritan A, Seno F (2006) Insight into the structure of amyloid fibrils from the analysis of globular proteins. PLoS computational biology 2: e170.
26. López De La Paz M, Serrano L (2004) Sequence determinants of amyloid fibril formation. Proceedings of the National Academy of Sciences of the United States of America 101: 87.

27. Fernandez-Escamilla A-M, Rousseau F, Schymkowitz J, Serrano L (2004) Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. Nature biotechnology 22: 1302–1306.
28. Galzitskaya OV, Garbuzynskiy SO, Lobanov MY (2006) Prediction of amyloidogenic and disordered regions in protein chains. PLoS computational biology 2: e177.
29. Thompson MJ, Sievers SA, Karanicolas J, Ivanova MI, Baker D, et al. (2006) The 3D profile method for identifying fibril-forming segments of proteins. Proceedings of the National Academy of Sciences of the United States of America 103: 4074–4078.
30. Belli M, Ramazzotti M, Chiti F (2011) Prediction of amyloid aggregation in vivo. EMBO Rep 12: 657–663.
31. Dou Y, Baisnée P-F, Pollastri G, Pécout Y, Nowick J, et al. (2004) ICBS: a database of interactions between protein chains mediated by β-sheet formation. Bioinformatics 20: 2767–2777.
32. Feverati G, Achoch M, Zrimi J, Vuillon L, Lesieur C (2012) Beta-Strand Interfaces of Non-Dimeric Protein Oligomers Are Characterized by Scattered Charged Residue Patterns. PloS one 7: e32558.
33. Lopez De La Paz M, Goldie K, Zurdo J, Lacroix E, Dobson CM, et al. (2002) De novo designed peptide-based amyloid fibrils. Proc Natl Acad Sci U S A 99: 16052–16057.
34. Richardson JS, Richardson DC (2002) Natural -sheet proteins use negative design to avoid edge-to-edge aggregation. Proceedings of the National Academy of Sciences 99: 2754.
35. Feverati G, Lesieur C (2010) Oligomeric interfaces under the lens: gemini. PloS one 5: e9897.
36. Feverati G, Lesieur C, Vuillon L.SYMMETRIZATION: RANKING AND CLUSTERING IN PROTEIN INTERFACES. In: Michel Deza MP, Krassimir Markov editor; 2012; Bulgeria. pp. p 133–146.
37. Faure G, Bornot A, de Brevern AG (2008) Protein contacts, inter-residue interactions and side-chain modelling. Biochimie 90: 626–639.
38. Andreeva A, Murzin AG (2011) Structural classification of proteins and structural genomics: new insights into protein folding and evolution. Acta Crystallogr Sect F Struct Biol Cryst Commun 66: 1190–1197.
39. Andreeva A, Howorth D, Chandonia JM, Brenner SE, Hubbard TJ, et al. (2008) Data growth and its impact on the SCOP database: new developments. Nucleic Acids Res 36: D419–425.
40. Murzin AG, Brenner SE, Hubbard T, Chothia C (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 247: 536–540.
41. Ma B, Nussinov R (2000) Molecular dynamics simulations of a [beta]-hairpin fragment of protein G: balance between side-chain and backbone forces1. Journal of molecular biology 296: 1091–1104.
42. Lo Conte L, Chothia C, Janin J (1999) The atomic structure of protein-protein recognition sites. J Mol Biol 285: 2177–2198.
43. Minor Jr DL, Kim P (1994) Measurement of the b-sheet-forming propensities of amino acids. Nature 367: 660–663.
44. Smith CK, Regan L (1997) Construction and design of β-sheets. Acc Chem Res 30: 153–161.
45. Fooks H, Martin A, Woolfson D, Sessions R, Hutchinson E (2006) Amino acid pairing preferences in parallel β-sheets in proteins. Journal of molecular biology 356: 32–44.
46. FarzadFard F, Gharaei N, Pezeshk H, Marashi SA (2008) [beta]-Sheet capping: Signals that initiate and terminate [beta]-sheet formation. Journal of structural biology 161: 101–110.
47. Adessi C, Soto C (2002) Beta-sheet breaker strategy for the treatment of Alzheimer's disease. Drug development research 56: 184–193.
48. Amaral LA, Scala A, Barthelemy M, Stanley HE (2000) Classes of small-world networks. Proc Natl Acad Sci U S A 97: 11149–11152.
49. Smith JM, Jang Y, Kim MK (2007) Steiner minimal trees, twist angles, and the protein folding problem. PROTEINS: Structure, Function, and Bioinformatics 66: 889–902.
50. Levin KB, Dym O, Albeck S, Magdassi S, Keeble AH, et al. (2009) Following evolutionary paths to protein-protein interactions with high affinity and selectivity. Nature structural & molecular biology 16: 1049–1055.
51. Greene LH, Higman VA (2003) Uncovering network systems within protein structures. J Mol Biol 334: 781–791.
52. Albert R, Jeong H, Barabasi AL (2000) Error and attack tolerance of complex networks. Nature 406: 378–382.
53. Albert R, Barabasi AL (2000) Topology of evolving networks: local events and universality. Phys Rev Lett 85: 5234–5237.
54. Higashimoto Y, Asanomi Y, Takakusagi S, Lewis MS, Uosaki K, et al. (2006) Unfolding, aggregation, and amyloid formation by the tetramerization domain from mutant p53 associated with lung cancer. Biochemistry 45: 1608–1619.
55. Barabasi AL, Oltvai ZN (2004) Network biology: understanding the cell's functional organization. Nat Rev Genet 5: 101–113.
56. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat T, et al. (2000) The protein data bank. Nucleic acids research 28: 235–242.
57. Tsai CJ, Lin SL, Wolfson HJ, Nussinov R (1996) A dataset of protein-protein interfaces generated with a sequence-order-independent comparison technique. J Mol Biol 260: 604–620.
58. Guerois R, Nielsen JE, Serrano L (2002) Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. Journal of molecular biology 320: 369–387.
59. Schymkowitz J, Borg J, Stricher F, Nys R, Rousseau F, et al. (2005) The FoldX web server: an online force field. Nucleic Acids Research 33: W382–W388.

Chapitre 8 : Intermolecular β-Strand Networks Avoid Hub Residues and Favor Low Interconnectedness: A Potential Protection Mechanism against Chain Dissociation upon Mutation

Chapitre 9: Étude de distributions de degré pour 40 protéines

Mon travail de thèse consiste à chercher les paramètres qui définissent les chemins de communication entre résidus à l'intérieur d'une structure, et ce dans le but de déterminer s'il existe des lignes de faiblesse dans le réseau pouvant expliquer la tendance de certains oligomères à changer de conformations. Il s'agit d'étudier l'architecture du réseau afin de comprendre les voies de communication du réseau. La distribution des degrés est une mesure classique de l'architecture d'un réseau, je l'ai donc mesurée sur les interfaces des 40 protéines oligomériques préalablement étudiées (Chapitre 7).

L'analyse des 40 protéines a pour objectif de distinguer le type de réseau associé à chacune d'entre elles, soit un réseau aléatoire (distribution de Poisson), soit un réseau sans échelle (distribution loi de Puissance) ou encore un réseau hiérarchique (Distribution exponentielle). Les résultats présentent la fréquence de la distribution de degré en fonction des degrés avec le coefficient de détermination pour les 40 protéines. Le calcul du coefficient R^2 et du degré moyen <k> des protéines considérées sont décrites pas des réseaux avec échelle soit suivant une loi exponentielle soit suivant une loi linéaire et rarement par des réseaux sans échelle.

Les architectures de réseaux étudiées indiquent que les interfaces des 40 protéines ne sont pas organisées en réseaux dont les faiblesses seraient liées aux degrés des hots pots. Un objectif clé de la recherche biologique est un catalogue systématique de toutes les molécules et leurs interactions au sein d'un réseau complexe.

9.1 Cadre du problème

9.1.1 Réseau protéique

Les interactions dans les protéines sont souvent représentées sous la forme d'un graphe. Le graphe est constitué de points représentant des résidus individuels reliés entre eux. Les connexions y représentent des interactions détectées expérimentalement ou prédites in-silico. Le réseau est caractérise par une distribution inégale du nombre de connexions parmi les acides aminés.

Dans l'analyse de la topologie des réseaux d'interactions, Vidal et Barabasi observent une propriété biologique émergeant de la structure même du réseau. Cette propriété est l'étonnante résistance aux mutations. La redondance génétique peut contribuer à une résistance [121]. Pourtant, nombreuses sont les protéines dont les gènes ne sont pas dupliqués. Les pertes d'activité de certaines de ces protéines devraient affecter le phénotype de

l'organisme. Cependant, on peut supposer que l'intégration de ces protéines déficientes à un vaste réseau d'interactions pourrait permettre la compensation de la perte de leurs activités [122]. Les mécanismes de résistance existent en biologie, par exemple dans les interactomes. Mon travail consiste à étudier des réseaux biologiques connus, réseaux d'acide aminés.

La caractéristique la plus élémentaire d'un nœud est son degré (sa connectivité), k, qui nous dit combien de liens le nœud a avec les autres nœuds. Un réseau non orienté avec N nœuds et liens L est caractérisé par un degré moyen $\langle k \rangle = 2L / N$. On appelle degré du sommet v (v est le nœud), et on note d(v), le nombre d'arêtes incidentes à ce sommet (arête égale nombre de liens).

9.1.2 La distribution de degré des trois types de graphe (voir chapitre 2)

La distribution des degrés, P (*k*), donne la probabilité qu'un nœud sélectionné ait exactement *k* liens. P (k) est obtenu en comptant le nombre de nœuds N (k) avec k = 1, 2 ... liens et en divisant par la somme nombre de nœuds N. La distribution des degrés nous permet de distinguer entre différentes catégories (architectures) de réseaux. Il existe trois modèles de graphes qui ont aidé à comprendre la distribution de degré dans les réseaux. La distribution de degré d'un graphe dépend de type de graphe :

9.1.2.1 Graphe aléatoires

Les degrés de nœuds suivent une distribution de Poisson, ce qui veut dire que la plupart des nœuds ont à peu près le même nombre de liens (proche du moyen degré <k>). La queue (de région de haute k) de la distribution de degré P (k) décroît de façon exponentielle, ce qui indique que les nœuds qui dévient considérablement de la moyenne sont extrêmement rares. La longueur de trajet moyenne entre n'importe quel nœud du réseau est proportionnelle au logarithme de la taille du réseau, ce qui indique qu'il est caractérisé par la propriété petit-monde.

9.1.2.2 Réseaux sans échelle

Les réseaux sans échelle sont caractérisés par des degrés suivant une distribution de loi de puissance; la probabilité qu'un nœud ait des liens k suit $P(k) \sim k^{-\gamma}$, où γ est l'exposant de degré. La probabilité qu'un nœud soit fortement connecté est statistiquement plus significative que dans un graphe aléatoire, les propriétés du réseau étant souvent déterminées par un nombre relativement restreint de nœuds hautement connectés qui sont connus comme centres (Hub). La plupart des réseaux sans échelles biologiques et non biologiques ont des exposants entre 2 et 3.

9.1.2.3 Réseaux hiérarchiques

Le modèle de réseau hiérarchique intègre de façon transparente une topologie sans échelle avec une structure modulaire inhérente en générant un réseau qui a une distribution de degré de la loi de puissance avec un exposant de degré $\gamma = 1 + \ln 4 / \ln 3 = 2.26$. Une architecture hiérarchique implique que des nœuds faiblement connectés fassent partie de zones très rapprochées (quartiers), la communication entre les différents quartiers très rapprochés étant maintenue par quelques hubs.

9.2 Protocole

Pour étudier les distributions de degrés des 40 protéines, j'ai utilisé Spectral-Pro pour générer les matrices d'adjacence et les 40 réseaux, et j'ai calculé les poids (interactions entre atomes) et les degrés (interactions entre acides aminés) pour chacun des réseaux. Pour cela, j'ai exécuté les étapes suivantes :

- Générer les fichiers PDB d'une base de données composée par 40 protéines.
- Exécuter le programme Spectral-Pro
- Analyser les résultats obtenus

Le calcul du nombre d'interactions entre les atomes et/ou entre les acides aminés nous permet d'analyser et d'interpréter l'existence du phénomène de redondance à partir du degré de distribution.

J'ai calculé ensuite la fréquence et le degré moyen de tous les cas étudiés.

9.3 Résultats

Une principale caractéristique d'un graphe est que la distribution de ses degrés. Il s'agit donc ici de voir sur un échantillon de 40 protéines, s'il existe une architecture unique de réseau d'acides aminés en interaction à partir de laquelle comprendre les chemins de communication privilégiés et les faiblesses du réseau.

Les résultats obtenus par le programme Spectral-Pro, les poids et degrés pour les 40 protéines étudiées, ont montré que les 40 cas se comportent de façon similaire en calculant la fréquence et la moyenne du degré de distribution. Je présente dans le tableau 9.1 suivant la fréquence du degré en fonction de degré.

1EEI		
Degré	**nombre de contacts**	**fréquence**
1	68	0,24
2	45	0,16
3	61	0,22
4	53	0,19
5	36	0,13
6	7	0,02
7	5	0,02
8	2	0,01
9	6	0,02

1SAC		
Degré	**nombre de contacts**	**fréquence**
1	84	0,41
2	31	0,15
3	41	0,20
4	18	0,09
5	16	0,08
6	10	0,05
7	1	0,00
8	2	0,01
9	2	0,01

2BAZ		
Degré	**nombre de contacts**	**fréquence**
1	55	0,3
2	22	0,12
3	41	0,22
4	39	0,21
5	11	0,06
6	5	0,03
7	8	0,04
8	0	0
9	3	0,02

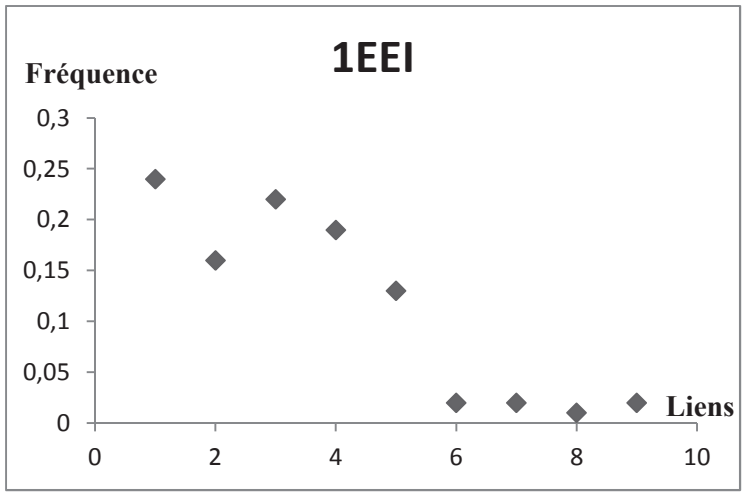
2V9U		
Degré	**nombre de contacts**	**fréquence**
1	115	0,32
2	82	0,23
3	61	0,17
4	57	0,16
5	14	0,04
6	9	0,02
7	16	0,04
8	0	0
9	8	0,02

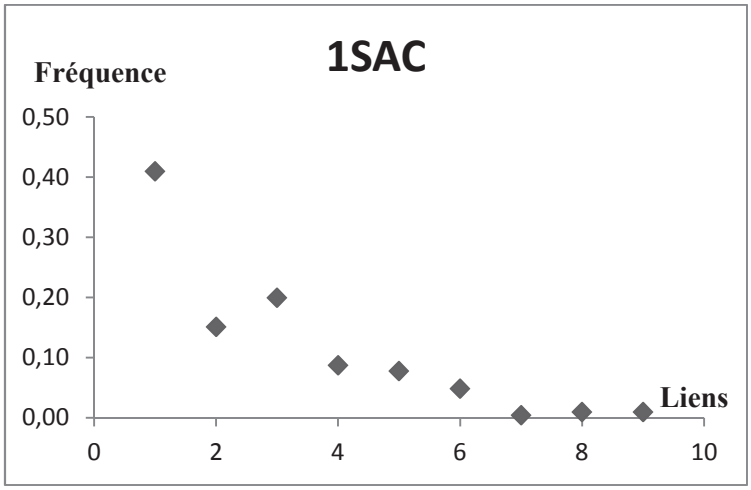
Tableaux 9.1 : La fréquence des quatre protéines parmi les 40 étudiés.

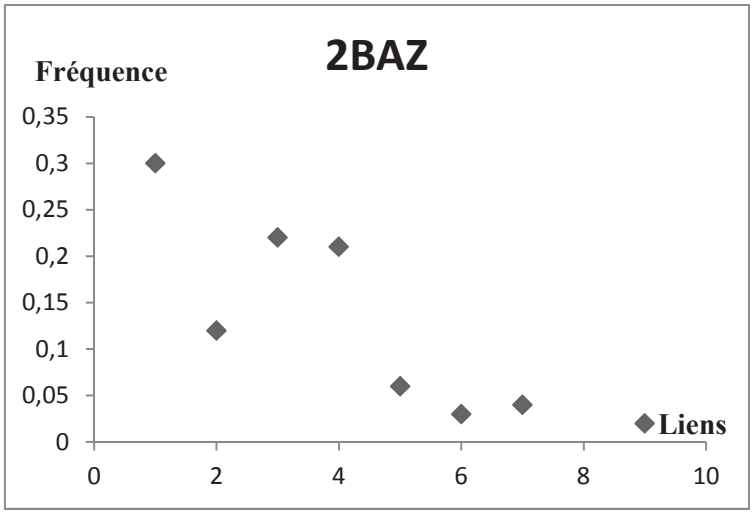
Parmi les 40 protéines étudiées, j'ai choisi quatre protéines ou les codes correspondants sont : 1EEI, 1SAC, 2BAZ et 2V9U. Les quatre exemples illustrent la distribution de degrés. La première colonne correspond au degré, le nombre de liens est présenté dans deuxième colonne et la troisième colonne est la fréquence de la distribution de degré pour chaque degré.

Les résultats obtenus par les quatre protéines montrent que le degré 1 à la plus haute fréquence et que la fréquence diminue avec l'augmentation du degré.

L'analyse des 40 protéines a pour objectif de distinguer le type de réseau associé a chacune d'entre elles, soit un réseau aléatoire (distribution de Poisson), soit un réseau sans échelle (distribution loi de Puissance) ou encore un réseau hiérarchique (Distribution exponentielle). Pour cela les figures suivantes présentent la fréquence de la distribution de degré en fonction des degrés avec le coefficient de détermination pour les quatre protéines choisies.


1EEI
Fréquence
0,3
0,25
0,2
0,15
0,1
0,05
0
0
2
4
6
8
10
Liens


1SAC
Fréquence
0,50
0,40
0,30
0,20
0,10
0,00
0
2
4
6
8
10
Liens


2BAZ
Fréquence
0,35
0,3
0,25
0,2
0,15
0,1
0,05
0
0
2
4
6
8
10
Liens

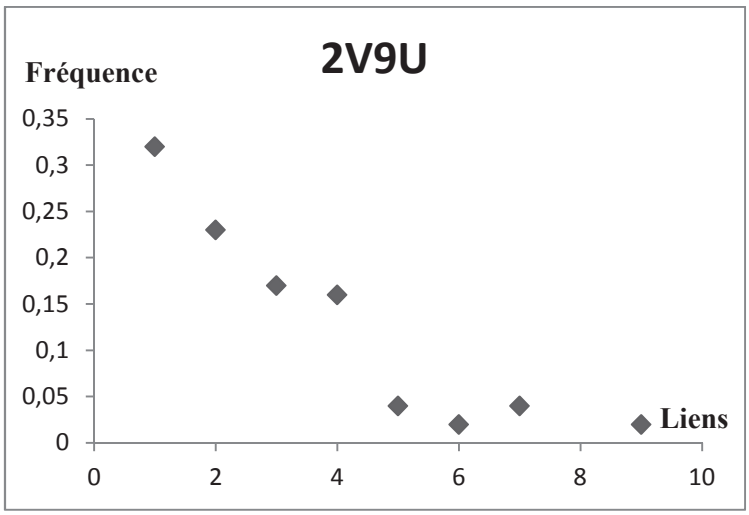


Figure 9.1 : les représentations des quatre protéines présentent la fréquence de la distribution de degré en fonction de liens.

On regardant le sous réseau de l'interface. J'ai obtenu les résultats suivants : les figures 9.1 présentent la distribution de fréquence en fonction du degré d'acide aminé. Le calcul des R^2 dans le tableau 9.2 (une distribution suivant une loi de puissance ou une exponentielle).

D'après les figures 9.1, on remarque que les quatre protéines (similaire pour les 40 cas) ont presque tous le même coefficient R^2. Calculant aussi le degré moyen <k> de chaque protéine (tableau 9.2), elles sont tous à un degré entre 2 et 3. Cette analyse a été faite pour les 36 autres cas et montre que les protéines considérées sont décrites pas des réseaux avec échelle soit suivant une loi exponentielle (50 % des cas) soit suivant une loi linéaire (37 % des cas) et rarement par des réseaux sans échelle (<15 % des cas). Le degré moyen autour de 2,9±0,1.

Protéines	power law (R^2)	exp (R^2)	random		linear y=ax+b		<l>	N	normal <l>		
			<k>	Ecart type	R^2	a			<l> ~$N^{1/2}$	diff<l>	>or<3
1B09	0,6	0,7	2,65	0,14	0,85	0,44	8	142	12	4	petit
1EEI	0,7	0,8	3,11	0,11	0,83	0,03	19	283	17	-2	normal
1EFI	0,78	0,87	3,11	0,11	0,82	0,03	17	317	18	1	normal
1FB1	0,35	0,53	2,71	0,10	0,78	0,05	18	240	15	-3	normal
1HI9	0,76	0,84	2,54	0,08	0,83	0,05	11	164	13	2	normal
1HX5	0,76	0,84	2,54	0,12	0,83	0,06	25	164	13	-12	grand
1J8D	0,73	0,85	3,14	0,14	0,77	0,28	24	220	15	-9	grand
1JN1	0,78	0,79	2,73	0,13	0,78	0,05	15	177	13	-2	normal
1L3A	0,82	0,9	2,96	0,12	0,82	0,03	22	289	17	-5	grand
1NQU	0,72	0,88	2,88	0,09	0,94	0,04	28	340	18	-10	grand
1PM4	0,86	0,82	2,91	0,23	0,77	0,04	14	79	9	-5	grand
1PVN	0,7	0,88	3,15	0,09	0,89	0,03	38	433	21	-17	grand
1SAC	0,77	0,86	2,55	0,12	0,73	0,04	9	205	14	5	petit
1SJN	0,75	0,9	3,11	0,13	0,88	0,03	16	237	15	-1	normal
1SNR	0,8	0,8	3,09	0,12	0,8	0,03	23	320	18	-5	grand
1TOA	0,71	0,86	2,72	0,20	0,79	0,04	11	69	8	-3	normal
1U1S	0,59	0,8	3,29	0,13	0,87	0,03	17	219	15	-2	normal
1WNR	0,57	0,67	3,03	0,15	0,75	0,04	25	148	12	-13	grand
1WUR	0,84	0,85	2,51	0,15	0,82	0,06	30	110	10	-20	grand
1Y13	0,76	0,93	2,68	0,14	0,97	0,05	10	127	11	1	normal
2A7R	0,8	0,94	2,96	0,12	0,84	0,03	29	293	17	-12	grand
2BAZ	0,71	0,82	2,97	0,14	0,7	0,03	19	184	14	-5	grand
2BCM	0,76	0,62	3,13	0,25	0,6	0,02	8	113	11	3	normal
2BHR	0,93	0,81	2,34	0,30	0,61	0,07	6	29	5	-1	normal
2BT9	0,94	0,9	2,34	0,14	0,8	0,05	9	127	11	2	normal
2GJV	0,73	0,83	2,72	0,17	0,65	0,04	11	100	10	-1	normal
2H5X	0,66	0,86	2,85	0,11	0,93	0,04	10	200	14	4	petit
2I9D	0,62	0,75	3,01	0,11	0,86	0,04	14	136	12	-2	normal
2JCA	0,68	0,81	3,07	0,19	0,7	0,03	13	110	10	-3	normal
2OJW	0,7	0,85	3,08	0,08	0,9	0,03	29	516	23	-6	grand
2P90	0,64	0,76	3,31	0,16	0,69	0,03	17	160	13	-4	grand
2RAQ	0,41	0,67	3,47	0,10	0,68	0,03	21	291	17	-4	grand
2RCF	0,67	0,86	3,05	0,12	0,83	0,03	20	232	15	-5	grand
2V9U	0,8	0,84	2,76	0,10	0,83	0,04	15	362	19	4	petit
2XSC	0,58	0,76	2,95	0,15	0,51	0,03	17	149	12	-5	grand
2Z9H	0,66	0,82	2,96	0,10	0,82	0,03	15	297	17	2	normal
3BF0	0,89	0,83	2,25	0,20	0,73	0,05	19	63	8	-11	grand

Tableau 9.2 : Résultats du calcul de coefficient R² et du degré moyen <k> pour les 40 protéines.

Les résultats montrent que l'écart entre les degrés est trop petit pour avoir une architecture sans échelle (loi de Puissance) avec des hubs qui contrôlent la communication [123]. Ceci est confirmé par les mesures de diamètres moyens (<*l*>) qui sont de tailles moyennes, c'est-à-dire autour de √N avec N le nombre de nœuds du réseau (Tableau 9.2, normal) ou grands, c'est-à-dire au-dessus de √N (Tableau 9.2, grand) et rarement petits, c'est-à-dire inferieur à √N (Tableau 9.2, petit). Un réseau petit monde a un diamètre moyen autour de Log N pour un réseau asymptotique, ce qui n'est pas le cas des réseaux considères ici qui ont des N petits. Les réseaux des interfaces de ces 40 cas ne communiquent donc probablement pas via des hubs et les hots pots de hauts degrés ne sont probablement pas plus sensibles aux mutations que les nœuds moins connectés. Il est intéressant d'observer des réseaux avec échelle car cela implique un coût de formation des liens et introduit un nombre de liens total constant et distribué sur des nœuds de degrés différents sur l'ensemble de ses interfaces. Dans les cas de distributions de degrés linéaires, la valeur moyenne de la pente est 0.04 ± 0.01 (le cas 1BO9 est omis), le faible écart-type supporte l'idée d'une invariance du nombre de liens pour former ce type d'interface β. Une analyse de clustering spectral appliqué au prototype de l'interface de CtxB indique que les hots pots d'une même région d'interface n'appartiennent pas tous aux mêmes clusters (**Annexe 5**). Un cluster se définissant en termes de connectivités (chapitre 2), ce résultat préliminaire renforce aussi l'idée que les interfaces se construisent sur un ensemble de propriétés de connectivités distribuées sur différents nœuds (hots pots). Ceci renforce l'idée que les interfaces sont construites pour répondre à une certaine invariance en termes de connectivité. Il est tentant de spéculer que cette constance du nombre de liens/connectivité fournit l'affinité nécessaire pour construire une interface stable.

Les résultats de cette étude des architectures de réseaux indiquent que les interfaces des protéines considérées ne sont pas organisées en réseaux dont les faiblesses seraient liées aux degrés des hots pots.

Chapitre 10: Article publie : Protein subunit association: NOT a social network

L'étude des réseaux des interfaces β (Chapitre 7, 8 et 9) suggère que la faiblesse des interfaces est liée aux degrés des hots spots. Le but du chapitre 10 est de tester cette hypothèse plus avant en regardant l'effet de mutations en fonction du degré du résidu muté afin de savoir si les nœuds responsables de la fragilité des interfaces sont les nœuds de hauts degrés ou les nœuds de bas degrés. Pour cette étude, un programme a été construit, appelé Spectral-Pro, programme qui mesure la connectivité des nœuds, c'est-à-dire le nombre de liens atomiques qu'ils font avec les atomes des acides aminés dans leur voisinage à une distance inférieure à 5Å. Spectral-pro applique une analyse spectrale sur chacune des composantes connexes individuelles ou il projette chaque paire d'atome sur un point dans l'espace.

L'analyse spectrale permet de regrouper les nœuds fortement interconnectés ensemble et faiblement connectés avec des nœuds extérieurs au groupe. J'ai mesuré le degré des nœuds avec Spectral-pro sur l'interface $CtxB_5$, ensuite j'ai muté K69N (degré 1) et R67N (degré 9) qui le plus haut du réseau. Fold-X a calculé les effets des mutations sur l'énergie d'interaction. Les résultats montrent que la mutation la mutation de K69N modifie plus l'énergie que la mutation de R67N. Donc le degré haut ne peut pas anticiper d'un effet plus grand que la mutation d'un bas degré. Ce résultat nous a mené à entreprendre l'étude de la mutation de tous les hot spots pour évaluer le rôle du degré sur la mutation.

Les résultats de cet article viennent compléter ceux de chapitre précédent, l'étude de la distribution de degré des 40 protéines, les approches réseaux sont l'outil pour analyser ces résultats.

Protein subunit association: NOT a social network*

Mounia Achoch†
LISTIC,University of Savoie, Annecy le Vieux, France

Giovanni Feverati‡
LAPTH UMR 5108, University of Savoie, CNRS, Annecy le Vieux, France

Laurent Vuillon§
LAMA UMR 5127, University of Savoie, CNRS, Le Bourget du Lac, France

Kave Salamatian¶
LISTIC, University of Savoie, Annecy le Vieux, France

Claire Lesieur‖
AGIM FRE 3405, University of Grenoble Alpes, CNRS, Grenoble, France

4th of April 2014

Abstract

Most proteins cannot function as single unit but associate subunits via the formation of protein interfaces, to be biologically active. How the amino acids involved in subunit association, so-called hot spots, regulate the formation of a protein interface is still an open question. Here, we show how network and graph theories can help addressing the role of hot spots. We built a MatLab code called SpectralPro which identifies hot spots and reconstructs the protein interface as a subnetwork of hot spots in interaction, with the hot spots as nodes and the bonds between hot spots as links. Using as a case study, the cholera toxin B pentamer (five subunits), we investigate if the degree of a node, namely the number of contacts of a hot spot, is important in the formation of an interface. The degree of a node is known to be important in many real networks. For example in social networks, hubs control the communication between most nodes and as such are vulnerable to changes. But our result shows that in the toxin interface sub-graph hub-like nodes are less vulnerable to change than single link node.

* Work supported by the Federation de recherche FR2914, MSFI, Modelization, Simulations, Fundamental interactions

† e-mail address: Mounia.Achoch@univ-savoie.fr
‡ e-mail address: feverati@free.fr
§ e-mail address: laurent.vuillon@univ-savoie.fr
¶ e-mail address: kave.salamatian@univ-savoie.fr
‖ e-mail address: claire.lesieur@agim.eu

1. Introduction

Proteins are biological entities made of a chain of amino acids bound to one another in a specific order, called the primary structure or the amino acid sequence of the protein. Based on the sequence and the environment, the protein acquires a tridimensional shape called tertiary structure (3D-structure), suitable for its biological function. The set of reactions leading to the functional 3D-structure is the folding of the protein. It involves the formation of bonds/interactions between atoms of the amino acids of a single chain. These interactions are called intramolecular amino acid interactions. There exist proteins which function as oligomers by associating several copies of the same chains (homo oligomers) or of different chains (hetero oligomers). The association of chains forms the quaternary structure (4D-structure) of the proteins. The zone of contact between two associated chains is called the protein interface. The protein interface involves the formation of interactions/bonds between atoms of the amino acids of adjacent chains. These interactions are called intermolecular amino acid interactions. Among the amino acids involved in intermolecular amino acid interactions, only a subset is important for the formation of the interface, those are called hot spots [1].

Some protein oligomers are involved in diseases as virulence factors, like the notorious cholera toxin responsible for the cholera disease [2]. Understanding and predicting how such proteins assemble into oligomers is essential for designing appropriate inhibitors capable of preventing their pathological assemblies. The design of such inhibitor entails to identify the hot spots and understand their role in the formation of an interface. There are numerous algorithms capable of identifying hot spots from the 3D structure of protein oligomers whose atomic coordinates are

available from the Protein Data Base (www.rcsb.org/pdb/). However, these algorithms do not provide means to understand how the hot spots orchestrate the formation of an interface. We propose to consider hot spots as nodes and bonds between hot spots as links, and to build a subgraph or a subnetwork of hot spots in interaction to model the interface. Sub graph because it describes only a local feature of the protein chain, namely the interface and not the entire chain, which would be a graph. The hot spots can be distinguished by network measures and we can look for correlation between the network's measures and the importance of the hot spots in terms of interface formation. A good overview of network measures can be found in [3]. Our case of study is the cholera toxin B subunit pentamer ($CtxB_5$) produced by *Vibrio cholera*. We have written a Matlab code that reasonably identifies the hot spots of the $CtxB_5$'s interface and builds a sub-graph of the toxin's interface based on a matrix of contacts. We look if the degree of the nodes, namely the number of contacts of the hotspots, has any relevance in terms of the formation of the toxin's interface.

2. Methods

SpectralPro. SpectralPro uses the Cartesian coordinates of the atoms of the 3D-structure of $CtxB_5$ as an input. These coordinates can be extracted from the PDB under the PDB code 1EEI. Each chain of the pentamer is considered as a set of points in the space whose positions are the Cartesian coordinates (x, y, z) of the atoms of the chain. The atoms of the chain 1 constitute the set 1 (S1), the atoms of the chain 2, the set 2 (S2) and the atoms of the chain 5, the set S5. SpectralPro calculates distances between every atom of S1 and every atom of the four other sets (interchain distances) but ignores the

distances between atoms of a single set (intrachain distances). It chooses for every atom the 10 closest atoms and among these, it selects the pairs of atoms distant of a maximum of 5 Angstrom. Every atom is involved in a certain number of pairs, namely it has a certain numbers of contacts. SpectralPro builds a N x N matrix with the selected intermolecular atoms as the nodes N and the elements of the matrix as their number of contacts. SpectralPro also builds a coarse-grained matrix where the atoms are replaced by their respective amino acids as nodes. A weightless matrix is produced where the elements of the matrix are one when the amino acids have at least one pair of atoms in contact and zero when they don't. The weightless matrix provides for every amino acid, its number of amino acid contacts.

Fold X. The effect of a local change (amino acid mutation) on the formation of the toxin interface is measured by generating a virtual single point mutation on the toxin PDB with Fold X and by calculating the free energies of interactions at the interface for the non mutated (wild-type) and the mutated proteins [4]. The difference between the two energies measures the effect of the mutation. The amino acid plays a role in the formation of the interface if its mutation leads to a non zero energy difference.

3. Results and discussion

The goal of the investigation is to develop an appropriate tool to reconstruct the $CtxB_5$ interface as a sub-graph of hot spots in interaction, analyze some graph properties to determine their relevancy in terms of the toxin assembly.

3.1. Identification of hotspots

The first step is to test if SpectralPro is capable of identifying hot spots. The details on how SpectralPro detects

amino acid in contact is described in the methods. Because SpectralPro reads the atoms following the amino acid sequence of the chain and selects the closest atoms, it retraces a good reading of the geometry of the two surfaces that make the interface compared to a selection based simply on a cut-off distance. The cut-of distance at 5 Ansgtrom applied subsequently allows to choose the bonds the most chemically probable. It is unlikely that every atom makes ten chemical bonds (ten closest atoms), but the ten links provide a density of interactions instead of evaluating an exact number of interactions. The idea is to obtain an estimate of a probability of interactions of the amino acids. The coarse-grained amino acid sub-graph is built on a square matrix having as rows and columns the amino acids, ordered according to their location along the sequence. The elements of the matrix at position i, j have a one entry if the i-th and j-th amino acids have at least one pair of atoms in interaction (weightless sub-graph).

The sub-graph of the atoms in interaction over the five interfaces of the pentamer has 1498 nodes and 2830 links. In other words, the sub graph is made of 1498 atoms with 2830 closest atoms. The coarse-grained sub-graph of the amino acids in interaction has 283 nodes and also 2830 links (weighted sub-graph). Thus on average every atom has two closest atoms located within 5 Angstrom distance and every amino acid has about five atoms involved in a pairwise interaction. If a single link is counted for every pair of amino acids, the (weightless) sub-graph has 283 nodes and 422 links. To have an idea of the order of magnitude of a protein interface sub-graph, it is interesting to compare with the world wide web which has 200 million nodes (webpages) and 1.5 billion links, links between two pages.

The amino acids selected as in interaction by SpectralPro are compared to the detection of hot spots by three

other available programs (not shown). SpectralPro identifies 283 amino acid contacts over 5 interfaces, with an average of 57 ±1 hot spots per chain. If we consider the set S5, namely the chain E, SpectralPro identifies 56 hot spots against 39, 57 and 54 for Gemini, PSIBASE and SCOWLP, respectively. Gemini detects hot spots by selecting the mutually closest atoms yielding a more stringent selection than SpectralPro and less hot spots identified [5]. All hot spots detected by Gemini are identified by SpectralPro. PSIBASE as SpectralPro calculates the Euclidean distance to determine pairs of interactions [6]. SpectralPro identifies all the hot spots identified by PSIBASE expect three, making about 5 % false negative. Only one amino acid detected by SpectralPro is not detected by PSIBASE, making less than 2 % false positive. On average in PSIBASE, every hot spot has 5 atoms involved in a pairwise interaction as observed for SpectralPro. SpectralPro identifies all the hot spots identified by SCOWLP expect one, making less than 2 % false negative. There are three amino acids detected by SpectralPro but not by SCOWLP, making about 5 % false positive. SCOWLP identifies pairwise interactions using Eucledian distances and shape-based algorithms [7]. Globally the amino acids selected as hot spots by SpectralPro are consistent with those identified by other programs, supporting that SpetralPro detects hot spots reasonably.

3.2. The degree measure

On a previous study on a large dataset of 1048 interfaces involving the interactions between two beta -strands, we had measured the degree of the nodes of the sub-graph interfaces and looked at the degree distributions [8]. The sub-graphs were built with a different algorithm, called Gemini which selects only a framework of interactions, as mentioned above. The result indicates an exponential de-

gree distribution, no hubs and many nodes with one to three contacts. We have determined statistically that the only amino acids with more than three contacts are R, Y, L and W.

Now we look whether this result is confirmed using SpectralPro which sets less stringency on the selection of hot spots and the number of contacts. The average number of contacts $\overline{k}$ over the five $CtxB_5$ interfaces is 3.1 $\pm$1.8. Thus even with SpectralPro, the average number of contacts per residues remains around three.

The degree distribution *P(k)* is the number of hot spots with k degree plotted against the degree k. *P(k)* is calculated for each of the five interfaces of $CtxB_5$ and the average degree distribution and standard deviation is plotted against the degree (Figure 1).

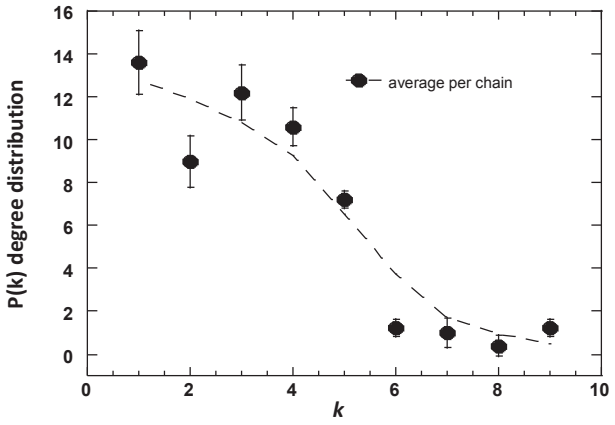


Figure 1: Degree distribution

P(k) for the sub-graph of the $CtxB_5$ interface follows

a bell like shape which corresponds to a random network with no hubs but nodes with few links. Again this confirms the observation made on the dataset using Gemini that interface subnetworks do not follow power law degree distribution and have no hubs.

At most the hot spots have 9 contacts, and there are only two such nodes, the amino acid arginine 67 (Arg67) and the leucine 31 (Leu31). Thus the bigger ratio between the highest and lowest degree in the sub-graph is 9. On a subgraph of the WWW of 325 729 nodes, which follows a power law degree distribution, the average $\overline{k}$ is 5.46, the ratio between the lowest and highest node degree is 10000. So the hot spots with 9 contacts might be better referred to as hub-like rather than hub. Interestingly, in comparison the average degrees $\overline{k}$ of the two networks appear rather similar, illustrating the difficulty in interpreting average $\overline{k}$ values for different types of degree distribution. This is discussed in [9].

3.3. Influential nodes

We then explore if the degree of the nodes is any relevant to the formation of the toxin interface. For this purpose, a hot spot with a single contact, lysine 69 (Lys69) and a hub-like hot spot, Arg67 are virtually mutated to an asparagine (Asp, N) using Fold X [4]. The free energy of interaction at the interface is calculated for the mutant and the wild type (WT) proteins. The effect of the mutation is measured as the difference between the wild type and mutant free energies of interaction at the interface. Differences not equals to zero indicate that the mutated hot spot is involved in the formation of the interface. Asparagine is chosen because it has "average" amino acid properties, so if a mutation has no effect on the free energy of interaction, it indicates that the mutated hot spot has average property for the formation of the interface and is plastic to mutation. If a

mutation has an effect, the mutated hot spot must have an involvement in the formation of the interface above average, this hot spot can be considered more influential for the formation of the interface and less plastic to mutation. The WT, Lys69Asp and Arg67Asp free energies of interaction are -13,35; -19, 65 and -16, 65 kcal/mol, respectively, as determined by Fold X. This shows that the hot spots are not equally important for the formation of the interface, suggesting their different roles. The free energy of the interface has decreased by a factor of 0.4 and 0.2 upon mutation of the Lys69 and Arg67, respectively. The largest mutational effect on the free energy is for the Lys69Asp mutant over mutation of all other amino acids of the toxin (not shown). Thus the mutation of the single link hot spot Lys69 has more effect on the interface than the mutation of the hub-like Arg67. Thus in contrast to social networks and other real networks, in the sub-graph of the toxin interface, the influence of a node is not directly linked to its degree. More precisely, hub-like residues are not more vulnerable to change, namely mutation, than single link node.

4. Conclusion

In conclusion, we can say that protein interface subnetworks have very different scales compared to other real networks, much less links, lower ratio high degree/low degrees, no hub and behave rather like a random network. Thus to infer "biological rules", such as the mechanism of assembly or the formation of interfaces, one cannot simply use the network measures that regulate other real networks (www or social network). Intuitively, we could have expected that hub-like hot spots would have been the most influential for the formation of the interface and highly susceptible to mutation as demonstrated for other real net-

works [10], but that is not the case . Here the result shows that connected does not imply influential in the case of protein interface networks. It remains to be established what makes a node influential if not its degree and to analyze the effect of the mutation on the network.

References

[1] Clackson T, Wells JA, *Science* 267(5196):383-6 (1995)

[2] Hirst TR, *J. Moss BI, M. vaughan and A. t. Tu, editor. New York: M. Dekker* 123-84 (1995)

[3] Barabasi A.L, Oltvai Z.N, *Nature reviews Genetics* Feb;5(2):101-13 (2004)

[4] Guerois R, Nielsen JE, Serrano L, *Journal of molecular biology* 320(2):369-87 (2002)

[5] Feverati G, Lesieur C, *PloS one* 5(3):e9897 (2010)

[6] Gong S, Yoon G, Jang I, Bolser D, Dafas P, Schroeder M, et al, *Bioinformatics* May 15;21(10):2541-3 (2005)

[7] Teyra J, Doms A, Schroeder M, Pisabarro MT, *BMC Bioinformatics* 7:104 (2006)

[8] Feverati G, Achoch M, Vuillon L, Lesieur C, *PloS one* in press (2014)

[9] Newman ME, Strogatz SH, Watts DJ, *Phys Rev E Stat Nonlin Soft Matter Phys* Aug;64(2 Pt 2):026118 (2001)

[10] Albert R, Jeong H, Barabasi AL, *Nature* Jul 27;406(6794):378-82 (2000)

Chapitre 11: Article en révision (PCCP) « Protein structural robustness to mutations: an in silico investigation»

L'étude de la fragilité des interfaces vis-à-vis de mutations (Chapitre 9 et 10) s'est poursuivie par un travail considérant la mutation de tous les hots spots de l'interface de $CtxB_5$ pour comprendre quel paramètre sous-tend la faiblesse. L'objectif est d'étudier les changements structuraux engendrés par les mutations individuelles de ses acides aminés, et de les considérer en termes de robustesse, adaptation et évolution.

La $CtxB_5$ est notre prototype d'étude ou les acides aminés qui composent l'interface de la toxine ont été étudiés. Après une mutation de chaque acide aminé par l'asparagine des 103 positions dans la chaîne protéique, j'ai généré les graphes des interfaces (pas seulement l'interface β) en utilisant les deux programmes GEMINI et Spectral-Pro. Ensuite j'ai fait des mesures de réseaux : degré local, degré global, poids local poids global et j'ai essayé de corréler ces changements de réseaux aux changements d'énergie d'interaction (stabilité de l'interface). Je n'ai pas réussi. Alors j'ai regardé les variations dans les hots pots détectés par Fold-X après mutations et je n'ai pas trouvé de différences. Donc malgré des changements de réseaux même important, Fold-X détecte toujours la même liste de hots pots mais il n'est pas possible de corréler les données d'énergie avec les meures de réseaux.

Le changement structural produit par les mutations ne dépend pas seulement du degré local d'un acide aminé, et l'impact fonctionnel d'une mutation ne se rapportent pas uniquement à la quantité des changements structurels. J'ai ensuite essayé de comprendre pourquoi certains acides aminés étaient modifiés et pas d'autres par divers calculs. Par exemple, j'ai compté les changements par paire au lieu de les compter par résidu, afin de trouver ou non le chemin de propagation suivant un flow. Ou encore j'ai compte les fractions de changements par résidu (changement du nombre lien après mutation divisé par nombre de liens initial). Mais à ce jour je n'ai pas pu mettre en évidence ce qui induit le changement. En dehors de la piste des backups.

Les qualités des protéines telles que la robustesse et de l'adaptabilité à des perturbations telles que des mutations sont étudiées dans cet article. L'impact structural de mutation est une enquête indépendante de l'impact fonctionnel. Les changements structurels peuvent se propager à partir du site de mutation de résidus beaucoup plus loin que les échelles typiques des interactions chimiques, à la suite d'un mécanisme de cascade. Ceci peut déclencher des changements dramatiques ou subtils, selon les maladies, ou de nouvelles fonctions. La robustesse est renforcée par les changements qui produisent des structures

alternatives, compatibles avec le point de vue que les protéines sont des objets dynamiques qui remplissent leurs fonctions à partir d'un ensemble de conformations. Les mécanismes sous-jacents et les effets structurels de mutations plus généralement sont étudiés en développant un algorithme appelé AminoAcidRank pour quantifier les changements structurels associés à des mutations, et aider à déterminer l'interaction (non additifs) effets.

PCCP

Protein structural robustness to mutations: an in silico investigation

Journal:	*Physical Chemistry Chemical Physics*
Manuscript ID	CP-ART-10-2015-006091.R1
Article Type:	Paper
Date Submitted by the Author:	n/a

Complete List of Authors:	Achoch, Mounia; University of Savoie, LISTIC Gilardi, Rodrigo; University of Savoie, LAMA Wymant, Chris; Imperial College London, Department of Infectious Disease Epidemiology Feverati, Giovanni; CNRS-UDS, MSIF Salamatian, Kave; University of Savoie, LISTIC Vuillon, Laurent; University of Savoie, LAMA Lesieur, Claire; CNRS, AMPERE

PCCP Guidelines for Referees

Physical Chemistry Chemical Physics **(*PCCP*) is a high quality journal with a large international readership from many communities**

Only very important, insightful and high-quality work should be recommended for publication in *PCCP*.

To be accepted in *PCCP* - a manuscript must report:

- Very high quality, reproducible new work
- **Important new physical insights** of significant general interest
- A novel, stand-alone contribution

Routine or incremental work should not be recommended for publication. Purely synthetic work is not suitable for PCCP

If you rate the article as 'routine' yet recommend acceptance, please give specific reasons in your report.

Less than 50% of articles sent for peer review are recommended for publication in *PCCP*. The current PCCP Impact Factor is **4.5**.

PCCP is proud to be a leading journal. We thank you very much for your help in evaluating this manuscript. Your advice as a referee is greatly appreciated.

With our best wishes,

Anna Simpson (pccp@rsc.org)
Managing Editor, PCCP

Prof Seong Keun Kim
Chair, PCCP Editorial Board

General Guidance (For further details, see the RSC's Refereeing Procedure and Policy)

Referees have the responsibility to treat the manuscript as confidential. Please be aware of our **Ethical Guidelines** which contain full information on the responsibilities of referees and authors.

When preparing your report, please:
• Comment on the originality, importance, impact and scientific reliability of the work;
• State clearly whether you would like to see the paper accepted or rejected and give detailed comments (with references) that will both help the Editor to make a decision on the paper and the authors to improve it;

Please inform the Editor if:
• There is a conflict of interest;
• There is a significant part of the work which you cannot referee with confidence;
• If the work, or a significant part of the work, has previously been published, including online publication, or if the work represents part of an unduly fragmented investigation.

When submitting your report, please:
• Provide your report rapidly and within the specified deadline, or inform the Editor immediately if you cannot do so. We welcome suggestions of alternative referees.

PCCP

ARTICLE

Protein structural robustness to mutations: an *in silico* investigation

Received 00th January 20xx,
Accepted 00th January 20xx

DOI: 10.1039/x0xx00000x

www.rsc.org/

Mounia Achoch[a], Rodrigo Dorantes-Gilardi[b], Chris Wymant[c], Giovanni Feverati[d], Kave Salamatian[a], Laurent Vuillon[b] and Claire Lesieur[e†]

Proteins possess qualities of robustness and adaptability to perturbations such as mutations, but occasionally fail to withstand them, resulting in loss of function. Here the structural impact of mutations is investigated independently of the functional impact. Primarily, we aim at understanding the mechanisms of structural robustness, pre-requisite for functional integrity. The structural changes due to mutations propagate from the site of mutation to residues much more distant than typical scales of chemical interactions, following a cascade mechanism. This can trigger dramatic changes or subtle ones, consistent with a loss of function and disease, or the emergence of new functions. Robustness is enhanced by changes producing alternative structures, in good agreement with the view that proteins are dynamic objects fulfilling their functions from a set of conformations. This result, robust alternative structures, is also coherent with epistasis or rescue mutations, more generally with non-additive mutational effects and compensatory mutations. To achieve this study, we have developed the first algorithm, referred to as Amino Acid Rank (AAR), which follows the structural changes associated with mutations from the site of the mutation to the entire protein structure and quantifies the changes so mutations can be ranked accordingly. Assessing the paths of changes opens the possibility to assume secondary mutations for compensatory mechanisms.

Introduction

How proteins sustain and adapt their biological functions, or fail to do so, is a complex question. The structure and function of proteins are defined by amino acid sequences which naturally vary upon genetic mutations. The robustness of proteins against mutations depends on the impact on the protein function of the structural changes arising from the mutations, changes which are not much investigated[1]. Proteins are strongly resistant to single amino acid mutations: most amino acids can be mutated without loss of function[2], i.e. such mutations are functionally neutral. Less frequently, with a frequency about 10^{-9} per site, mutations lead to the emergence of new functions (innovation) [3]. Alternatively, there are pathological mutations which lead to a loss of function. The present view of neutral mutations is that some are adaptive because their combination with other mutations drives functional evolution through non-additive effects (e.g. functional promiscuity or epistasis) [3]. Non-additive effects are also involved in rescue mechanisms, where the negative effect of pathological mutation is neutralized by a mutation at a second site[2, 4-6]. Generally, protein robustness, protein innovation and protein adaptation refer to the impact of mutations on the biological function of proteins.

On the other hand, the structural changes which are tolerated by a protein without jeopardizing the protein functionality (functional robustness or emergence of a new function) or those who on the contrary lead to loss of functions, are rarely looked into. Yet, even little understanding of the underlying structural changes would be instructive to address pathological mutations or help designing new enzymes. The gap between the studies on functional and structural robustness is due to several issues. To investigate functional robustness, a protein prototype is chosen, every individual amino acid is mutated and the function of the mutants is tested experimentally[7]. Likewise, studying structural robustness, namely maintenance of the structural integrity necessary for a biological function, implies to choose a protein prototype, mutate every individual amino acid, crystallize each mutant, solve each structure and compare the ones which share the same function. First, this is technically and financially challenging as well as time consuming. Second, the goal is to understand if a protein structure is built to bear mutational changes and if so, to investigate by what mechanisms. Thus an experimental approach is not appropriate because some mutations would fail to produce a structure but for reasons not necessarily related to structural robustness. A mutation might prevent folding and acquisition of a stable structure, but have no impact on the structural robustness. For instance, the B subunits of the pentamers of the cholera toxin and the heat labile enterotoxin maintain a pentamer at pH 5.0 but do not reassemble at this pH [8-11]. Also a mutation leading to a new structure and a new function might not easily be identified as such, experimentally. On the other hand, *in silico* mutations produce structural changes in

[a] Laboratoire d'informatique systèmes, traitement de l'information et de la connaissance (LISTIC), Université de Savoie, Annecy le Vieux, France
[b] Laboratoire de mathématiques (LAMA UMR 5127), Université Savoie Mont Blanc, CNRS, Le Bourget du Lac, France
[c] Medical Research Council Centre for Outbreak Analysis and Modelling, Department of Infectious Disease Epidemiology, Imperial College London, London, United Kingdom
[d] Federation de recherche Fr3405, Modelisation, Simulations, Interactions Fondamentales, Annecy-le-Vieux, France
[e] CNRS-UCBL, IXXI-ENS-Lyon, Laboratoire AMPERE, Lyon, France
† Corresponding author: claire.lesieur@ens-lyon.fr
See DOI: 10.1039/x0xx00000x

order to generate a stable structure. *In silico* methods cannot create a new structure or destroy a structure from a mutation, they produce a set of conformations close to the wild-type structure. This is a relevant framework to investigate the structural changes which underlie structural robustness as a general issue rather than having to restrict the study on specific mutations. The third issue is the lack of tools to measure and compare the effects of mutations on a structure, comparison needed to understand the mechanisms by which the protein structure bears the changes. There exist programs to compare global structure features (e.g. RMSD) and visualize structural differences[12-15]. But here it is about following changes from a local perturbation, the site of the mutation, to the entire protein structure.

To circumvent these difficulties, we have adopted the following strategy. We have worked on the atomic structure of the pentamer of the cholera toxin B subunit ($CtxB_5$) because it is a stable protein with an OB-fold, structure common to many other proteins with different sequences. We can therefore assume that the structure is naturally robust to mutational changes. We have generated a set of *in silico* mutations using Fold X, which produces structural changes maintaining a reliable structure[16]. Let us recall that the goal of the study is not to predict the effects of experimental mutations on a structure, but to have a set of mutations appropriate to explore structural robustness. The dataset is the individual mutation of all the amino acids which compose the toxin interface. To analyse the structural changes due to mutations, we have modelled the toxin interfaces as networks of amino acids in interaction such that the structural properties are compared through network comparison. The analysis of the networks helped us to build an ad hoc algorithm, called Amino Acid Rank (AAR) which takes into account all structural changes observed in the dataset, quantifies them and ranks the mutations accordingly.

Finally, we have analysed the results of AAR in terms of structural robustness. The results indicate that mutations generate structural changes at different scales (local or long range) in a cascade mechanism and independently of the local changes on the mutation site and of the nature of the mutation. Structural robustness relies not only on mutations producing no or little changes but also on mutations producing significant structural changes but generating redundant conformations, in good agreement with the recent definition of protein as an ensemble of conformations fulfilling one function. Thus, the redundancy produces alternative structues necessary for having conformations functionally distinct upon secondary mutations, consistently with "adaptive neutral mutations". An example of non-additive mutations is provided not in the context of emerging functions but as a correction mechanism of a cancer-related mutation reported in the tetrameric domain of the tumour suppressor p53. This error-correction mechanism is not conceivable if structural robustness is based only on a lack of structural changes upon mutation. The identification of a second site mutation capable of correcting default is possible because of the new algorithm AAR.

Methods

AminoAcidRank (AAR) algorithm. *Function SpetralPro.* The goal is to model a protein interface by a hotspots network. A protein interface is made of the amino acids of one chain which interact with the amino acids of adjacent chains. These amino acids are referred to as hotspots. To construct a hot-spot network, we first define its atomic network. Using the atomic coordinates from a PDB, all distances between atoms of one chain and atoms of adjacent chains are computed. Two atoms share a link if they are within 5 Å distance. Two hotspots share a link if they have at least one of their respective atoms within 5 Å distance from one another. It is convenient to represent the hot-spot network as its adjacency matrix A. If N is the number of hotspots in the protein, then A is the N × N matrix with value $a_{i,j}$ in row *i* and column *j* if *i* and *j* are connected by a link, and 0 otherwise. The weighted adjacency matrix W is defined by $w_{i,j}$, the weight of the link connecting *i* and *j,* that is the number of atomic links between amino acid *i* and amino acid *j*. The adjacency matrix A is defined by $a_{i,j}$ equals 1 if w_{ij} >0, otherwise a_{ij} equals to 0.

Function Arank. A mutated PDB is generated with Fold X introducing a single hotspot mutation of a residue at position r. The function SpectralPro is then applied on the mutated PDB. To compute the quantity of structural changes produced by the mutation, a N × N « difference matrix » D is defined as follows: $d_{i,j} = w_{i,j}^{mut} - w_{i,j}^{wt}$ where $d_{i,j}$ is the entry value of D at row *i* and column *j*, $w_{i,j}^{mut}$ is the weight of the mutated network at row *i* and column *j* and $w_{i,j}^{wt}$ is the weight of the wild type (WT) network at row *i* column *j*.

The structural changes produced by the mutation on the entire structure (Global changes, $arank_r$) are computed as the sum of the absolute value of all the entries of D (that is $\Sigma_{i,j} |d_{i,j}|$). The structural changes at the position of the mutation (local changes, $local_r$) are computed as the sum of the absolute values of all entries of D at row j (that is $\Sigma_j |d_{i,j}|$). The $arank_r$ values are used to rank mutations according to the amount of structural changes they produce.

Function Backup. This is to compute the redundancy of every link of the WT hotspot network. The backup links are sought within the local secondary structure around every hospot link based on the known hydrogen bonding of secondary structure. That is any (i,j) links located within a distance of 4 residues along the sequence on both chains of the considered hotspot link is computed as its backup link. Details are provided in the AAR pseudocode.

The AAR pseudo code is provided in the electronic supplementary information (ESI).

Fold X. Mutations were computed using the protein design tool of Fold X (version 3 beta) [16, 17]. Only the protein design function was used for mutagenesis using the PDB 1EEI as the wild-type (WT) structure. Details and run parameters are in the electronic supplementary information (ESI). Essentially the run parameters are chosen to minimize their impact on the network construction, to be applicable broadly on X-ray structures, and not to depend too strongly on a high quality structure. Here the qualities of the structures need to be at ~ 2.5 Å or above resolution.

Results and discussion

The aim is to investigate the structural changes that a protein may go through from individual mutations of its amino acids, still maintaining a stable structure. As a model of study, we use $CtxB_5$,

focusing on the amino acids that compose the toxin interface, so-called hotspots. A protein structure is built on atomic interactions between its amino acids, likewise for a protein interface. Thus to analyze the structural changes that take place in the toxin interface upon mutation, first intermolecular atomic interactions need to be established. The exact atomic interactions are intractable due to the large size of the system. Atomic interactions rely on chemical nature of atoms, distances between atoms and the atom environment (atomic neighbors). In order to take these parameters into account, the following procedure is undertaken (Methods). The distances between all atoms of one chain and all atoms of an adjacent chain, referred to as interatomic distances are calculated from the X-ray coordinates of $CtxB_5$ provided by the RCSB Protein Data Bank (PDB code 1EEI). All interatomic distances within 5 Å are considered as chemical interactions, without distinguishing the nature of the atoms (methods). This approximation is reasonable because every type of chemical interactions (van der Walls, electrostatic, hydrogen bonds, etc) between the atoms of amino acids carbon, oxygen, nitrogen, sulfur and hydrogen fall within a distance of less than 5 Å [18]. The chemical nature of atoms is not considered also because it is assumed that two atoms in the X-ray structure would not be close if they ought to chemically clash. They are either necessarily chemically compatible or their neighbors' shielding prevent them from clashing.

To each hotspot is associated a weight w_i equals to $\sum_j wi,j$, which is the total number of its links (intermolecular atomic distances within 5 Å, see methods). The pairs of atoms which are within 5 Å distance are coarse-grained to their respective amino acids in order to associate to each hotspot a number of amino acids in physical contacts (degree), number a_i equals to $\sum_j ai,j$. Because all the distances within 5 Å of every atom are considered, the algorithm intrinsically accounts for the neighbor atoms. The weigh and the degree can be considered as proxy of the probability of interactions of the amino acid, the higher the degree the more likely the amino acid is to have an interaction.

Survey of the structural changes

Our algorithm Amino Acid Rank (AAR) after establishing the amino acids and the interactions that composed the toxin interface with the above procedure, models the interface as a network of amino acids in intermolecular interactions (Methods, Function SpectralPro). The amino acids that have at least one intermolecular atomic distance within 5 Å are linked and referred to as hotspots. The $CtxB_5$ interface has 58 hotspots forming the nodes of the network, these are also recognized as hotspots by other programs available [19]. There are no histidine or cysteine hotspots.

We systematically mutate every hotspot one by one. In the current work we restrict ourselves to mutations to asparagine residue for simplicity, asparagine having average chemical and geometrical properties. For example it has a residue that is polar rather than hydrophobic or charged, and has an average number of atoms as compared to other amino acids. Mutations to other amino acids will be considered in future work.

In silico mutations are performed using Fold X (Methods) [16] to generate a mutated structure, from which a mutated toxin interface and a mutated network are produced by the AAR algorithm. To capture the structural changes associated with a mutation, AAR compares the networks after and before mutation and extracts all modified amino acid links (Methods, Function Arank). Mutations change the positions of atoms which modify the intermolecular atomic distances and so the nodes, degrees and weights of the network. To quantify the structural changes produced by a mutation at position *r* within the entire structure ($arank_r$), AAR sums the absolute values of the differences between the weights after and before mutation of all the nodes of the networks, the higher the $arank_r$ the larger the structural changes (Table 1). A change in weight means some atoms have become closer or further away, implying atomic interaction rearrangements. Depletion of an amino acid link means that the two hotspots have no more atoms within 5 Å distance. Addition of a new link means that the two hotspots have moved closer so they have atoms within 5 Å distance. These are amino acid link rearrangements. To qualitatively describe the mutations, a sphere of influence is defined as the number of modified amino acids by the mutation and by the distances between the site of the mutation and the modified residue the furthest from it (Table 1). Two distances are measured, geodesic and Euclidian. The geodesic distance is measured by the number of chemical links to be crossed to go from the site of mutation to the modified residue the furthest from it by the shortest path and the Euclidian distance is measured between the two residues in Ångström (Fig. 1). The spheres of influence of the fifty eight mutations are shown on their respective X-ray structures in Fig. S1 (see the electronic supplementary information, ESI), highlighting the broad diversity of structural changes in quantity and quality. The $arank_r$ values vary from 182 to 2, ten mutations have an $arank_r$ below the first quartile while fifteen have an $arank_r$ above the third quartile, and thus most mutations generate significant changes (Table 1). The changes involve side chain atoms only since the RMSD is zero for all mutations. No more than 10 % of the native interfacial contacts are loss upon mutations. On average the mutations modified eight hotspots; a quarter modifies only up to five hotspots and a quarter modifies more than eleven. Thirteen mutations out of fifty-eight produce only local perturbations, namely structural changes of residues in physical contact with the site of the mutation and so located within the chemical reach of the mutated residue (Euclidian and geodesic distances within 5 Å and 1, respectively). Forty-five mutations produce global changes, namely changes beyond physical contact and chemical reach of the mutated residue. Eighteen modify residues located at distances above 10 Å. The maximum long range modification is 17 Å. The mechanism of the long range modifications is chemically sound since the changes are going from hotspots chemically linked to hotspots chemically linked in a step-by-step manner as determined from the geodesic distances (Fig. 1). This cascade mechanism seems related to the secondary structure of the mutated residue since out of eighteen residues belonging to α-helices, seventeen produce a cascade (long range changes) upon mutation (95 %). Out of twenty-six which belong to a β-structure, thirteen produce a cascade (50 %) while out of fourteen which belong to a loop, twelve produce a cascade (86 %). This relation would need to be verified and further explored on a dataset. There are twelve mutations for which the changes do not go from hotspots to hotspots but go from the mutated residue to its intramolecular contacts, which subsequently modify their hotspots (Table 1, column Intra). It is still a step-by-step mechanism, but through intramolecular and intermolecular links. Thus the results highlight paths of changes between amino acids of the interface and amino acids outside it. Likewise, mutations of amino acids outside the interface are capable of modifying hotspots' degrees (work in progress). This is consistent with the mechanisms of protein assembly combining folding and association steps in a coordinated manner (for review see [20]). A step-by-step mechanism is described in other real networks as Peer-to-Peer mechanisms (P2P)[21].

As selected examples, the mutations K69N, A64N, L31N and I39N are considered in details because they allow covering the chemical and geometrical properties of amino acids (small, medium and large side chain, hydrophobic, charged and polar chemical nature). Their spheres of influence are shown in the X-ray structures of the respective mutants (Figure 2A). Large modifications are seen for the K69N and A64N mutants while fewer modifications take place for the mutants I39N and L31N. The mutations K69N and A64N are among the top disruptive ones with $arank_r$ values equal to 182 (first rank) and 112 (fifth rank), respectively (Table 1). This highlights that the extent of the structural changes cannot be inferred by the difference between the nature of the original and mutated residue since lysine is bigger and has more atoms than asparagine while alanine is smaller and has less atoms. This is further supported by the fact that the mutations of other lysine or alanine such as K34N and A102N have different AAR values (Table 1). To consolidate this point the spheres of influences shown in figure S1 are sorted by amino acid type, subsequently sorted by decreasing values of $arank_r$.

Now if the mutation K69N is compared to the mutation L31N, the latter has an $arank_r$value ten times lower than 182. Yet the residue L31 has a degree 9 and a weight 74, significantly higher than the degree and weight of the residue K69, 2 and 29, respectively. Like the nature of the residue, the degree or the weight does not condition the extent of the structural changes. This is further evidenced by plotting the $arank_r$ values against the weight of the original residue before mutation for the fifty eight mutations (Fig. 2B). The linear correlation is weak (Fig. 2B, $R^2 = 0.27$), indicating that mutation of an amino acid with a high weight does not systematically lead to large structural changes, and likewise mutation of an amino acid with a low weight does not necessarily lead to few structural changes.

The $arank_r$ values are then plotted against the local weight changes ($local_r$, weight differences on the mutated residue after and before mutation, see methods), and again a rather weak linear correlation is observed (Fig. 2C, $R^2 = 0.44$). This indicates that global changes are not proportional to local changes. Moreover, only some mutations have $arank_r$ values which fall on the straight-line of slope two implying local changes (Fig. 2C, red line). Most mutations have $arank_r$ values outside this line and so the produce global changes and involve cascades. If there are only local changes, that is weight changes on the mutated residue and nowhere else, then the global changes are twice the local changes because the global changes count the weight changes on the mutated node and on its endpoint nodes. This confirms that mutations produce changes at different scales as shown by the spheres of influence (Fig. S1). The absence of correlations between the $arank_r$values and the local weight before mutation or the local weight changes remains true even if the networks are built with cut offs 4 and 6 Å instead of 5 Å. Thus these properties are invariant within the experimental error of X-ray structures (~ 1Å). It is interesting to discuss the two AAR outliers, the mutation R67N and the mutation K69N because they have similar local and global changes (Table 1). What is different however is their fraction of local changes: R67N has lost 24 % of its interactions (24/101, ratio local weight difference to weight before mutation) while K69N has lost 77 % (23/30). The fraction of local changes does not correlate either with the global changes measured by AAR (not shown).

Structural robustness, fragility and adaptation

To assess whether the structure of a protein is built to bear mutational effects, we propose to consider the structural changes produced in the CtxB interface by the mutations and see if they are consistent with all known mutational effects: robustness, innovation, adaptation/rescue and pathology.

The first key point is that the mutations yield structural impact at different scales (Table 1, Fig. 2, Fig. S1). This means there is no *a priori* specific scale (e.g. 5 Å) at which structural changes can be detectable and it is necessary to measure them locally as well as globally. This is in good agreement with other studies showing both direct and indirect physical interactions in co-evolving residues [1]. Local structural changes, namely modification within the chemical reach of the site of the mutation is consistent with enzymatic innovation or adaptation which does not lead to a full reorganization of the global structure. Global structural changes are consistent with pathologies where a single mutation is enough to jeopardize a structure and consequently a function. Of course, this does not imply that enzymatic innovation and pathology occurs only via local and global changes, respectively. This all depends on the scale at which the function is regulated by the structure.

The scaling does not explain adaptation through epistasis, rescue mechanism, or compensatory mutations (non-additive effects). Let us consider the pre-requisite for such effects: a mutation at a site 1 with an effect 1 (Mutant 1) and a mutation at a site 2 with an effect 2 (Mutant 2). Non-additive effects mean the consequences of the combination of mutations 1 and 2 are different from the consequences of mutation 2 (or of mutation 1) individually. This implies that the structures of the mutant 1 (or of mutant 2) and of the wild-type are different, otherwise they would react similarly upon the secondary mutation (Fig. 3). In other words, a robust mutation that leads to a rescue mechanism or a compensatory effect upon a second site mutation necessarily has a structure distinct from the WT one. This suggests that functional robustness is built on mutations with no structural impact (neutral mutation) as well as on mutations producing distinct structural solutions functionally equivalent to the WT one (adaptive mutations). If true, this means among networks different from the WT one (i.e. $Arank_r \neq 0$), some should be WT-alternative and other should be dissimilar. To investigate this possibility, the four mutations K69N, A64N, L31N and I39N are considered again. The structural changes due to these mutations are schematized by networks before and after mutation on Fig. 4. Let us first consider the mutations K69N and A64N which both have significant structural changes, namely high $arank_r$ (Fig. 4A). The K69N mutation modifies the layout of the WT network substantially, since it reduces the atomic interactions between the region of interface composed of residues 63 to 67 of one chain and residues 73 and 65 of the adjacent chain, and simultaneously increases the atomic interactions between the residue 67 of one chain and the residues 27 to 37 on the adjacent chain. This is well-illustrated on the X-ray structures (Fig. 4A). Moreover, the mutation also depletes the only two weak ties of the WT network, namely the links (31, 50) and (63, 53) which connect two regions of interface otherwise unconnected.

On the contrary, the networks A64 and N64 have a similar layout (Fig. 4A). In fact, the N64 network appears like a WT alternative network with more amino acid links, but the same regions are connected. The K69N and A64N mutations well-illustrate the distinction between structural changes and alternative structural solutions. The mutations I39N and L31N have low $arank_r$ (14 and 18, respectively) but a similar result can be observed (Fig. 4B). Only the link (39, 8) is depleted in the I39N mutation, not modifying the network significantly since there are other linked residues in the

vicinity of the link (39, 8) (Fig. 4B). In contrast, the L31N, even though it also yields a single link depletion (31, 50), the mutated network is not equivalent to the WT one because it lacks the only link that was connecting the regions 50, 64-68, 88 and 96-98 through the intermolecular link (31, 50) (Fig. 4B). It is therefore important to acknowledge that structural changes large or small yield alternative networks or not. So, the quality of structural changes must also somehow be incorporated in order to anticipate the impact of a mutation. Because of the scaling issue and the cascade mechanism, establishing the appropriate measure for alternative networks to sort out robust (neutral and adaptive) and fragile mutations is complex and beyond the scope of the present work.

The obvious difference between the A64N and I39N alternative networks and the altered K69N and L31N networks is the redundancy of amino acid and atomic links in the formers. This is reminiscent of peer-to-peer networks, which are robust to perturbation because they have more links than necessary -'back up' links- such that depletion or addition of links is tolerated by generating several alternative networks [22]. To see if alternative structures and networks exist in proteins, we have measured backup amino acid links in the interface of $CtxB_5$. Two amino acid links (*i, j*) and (*i', j'*), which belong to the same secondary structural element, defined as the residues -*i*- and –*i'*- are four amino acids apart along the sequence and likewise for their respective -*j*- and -*j'*- residues, are considered to backup each other. This is because the integrity of the secondary structure relies on at least the amino acid links which participate to the hydrogen bonding. The maximum distance of four amino acids apart along the sequence corresponds to a helix turn (*i* +/-4) so backup links are counted within this range of distance along the backbone. Based on this definition of backup, AAR calculates the number of backup links for each link of the WT network (Methods, Function backup). Out of 92 links of amino acids, only the two weak ties have no backup. Eleven links have 1 to 3 backups, fifty-two have 4 to 13 backups and twenty-seven have more than fourteen backups. A backup network of the WT toxin interface is shown in Fig. 5, with the number of backups of each link described by a colour code. The network shows a non-uniform distribution of the number of backup per hotspots within the structure that may indicate fragile areas. This result supports the possibility of having neutral structural changes through addition and/or depletion of links producing alternative networks and structural robustness (Fig. 5). The backup for the residues K69, A64, L31 and I39 are 16, 41, 80 and 26, respectively. The mutations A64N and I39N which have a redundant network also have a higher backup than K69N. The L31N has a highest backup but the amino acid link (31, 50) has none. This illustrates the complexity in assessing robustness due to the scaling problem (robustness of a node, of a link or of a region/community). Nevertheless the results are encouraging to further explore the concept of backup as a measure of robustness and fragility.

WT alternative networks lay the ground for non-additive mutational effects because different atomic interactions would cope differently with secondary mutations. A mutation not tolerated in a WT network/structure might be tolerated in a mutated WT alternative network. We tested this possibility to further support a mechanism of robustness via alternative WT networks. The cancer-related mutation G334V reported for the tetrameric domain of the tumour suppressor p53, is used as a default mutation case[23]. The goal is to find a second site mutation which alone produces neutral structural changes and a WT alternative network but coupled with the G334V mutation prevents its structural damages, corroborating non-additive effects through alternative networks. The impact of the G334V mutation on the protein conformation is such that X-ray crystallography is inapplicable and there is no fiber structure available yet. The mutation G334V is generated *in silico* from the WT atomic structure (PDB 1SAK) using Fold X instead. The interface between chains D and B is analysed. The G334V mutation leads to a large amount of structural changes as the AAR is 286, there are side chain and backbone atom rearrangements since the RMSD is 0.03 Å. The sphere of influence reveals long range changes up to residues at geodesic distances five and Euclidian distance 15 Å from the residue 334 (Fig. 6). The structural changes go from the residue 334 up to the residue 324 on the N-terminal end and up to the residue 352 on the C-terminal end (Fig. 6A). The mutation does not change the degree of the residue 334 but it changes the degree of its intramolecular amino acid neighbours, residues 333 and 337, in a cascade mechanism (Fig 6). As a result, the residue 337 loses its pairing with the residues 345, 349 and 352, keeps its pairing only with the residue 348, reducing the connectivity within the interface region composed of the residues 345 to 352 and 337 to 341 (Fig. 6B). Moreover, the residue 333 also loses pairing with the residue 345 removing a link between the interface region composed of residues 330-334 and 325-328 and the interface region composed of the residues 337-341 and 345-352 (Fig. 6B). It is possible that the rigidity between these two regions loosen up after depletion of the link 345-333. The residue N345 is at the cross-road of the structural changes produced by the mutation G334V. We tested if a mutation at this position could reinforce the atomic interactions of the network such that it becomes robust to the G334V mutation. Again *in silico* mutations are performed using Fold X. The network of the single mutant N345D is similar to the WT network except for an increase of the weights (number of atomic interactions) of the links (345, 333), (345, 341), (337, 348) and (337, 349) and a decrease of the weight of the link (337, 345) (Figure 6B). The double mutant N345D+G334V has structural changes on half as many residues as the mutant G334V, it maintains both links (345, 333) and (345, 337) and its network looks like the WT one, apart from an additional link between the residue 333 and 352 found as well in the single mutant G334V (Fig. 6B). The small changes in the atomic interactions produced by the N345D prevent the residue 337 from moving away after the mutation of the residue 334 and prevent the loss of the link (333, 345). This is a non-additive mutational effect since the effects of the individual mutations differ from the effects of combined mutations; the effects of the G334V are lost when combined with the N345D mutation. This suggests that a second site mutation producing a compensatory effect is to be found among the residues modified by the first site mutation, namely it is on the sphere of influence of the first site mutation. This hypothesis is supported by the observation that on average, in the interface of $CtxB_5$, eight amino acids are modified by mutation and on average deleterious mutations can be compensated by nine mutations [1, 24].

Conclusions

The work investigates the mechanisms proteins use to resist structural changes upon mutations, as a groundwork to understand functional robustness. Assuming that all proteins bear mutations by similar mechanisms, a case of study is a good model of investigation. The first challenge is to elaborate a set of mutations producing structural perturbations still maintaining a viable structure to look at. The solution proposed is to mutate *in silico*

every amino acid of the interface of the B subunit pentamer of the cholera toxin and to monitor structural changes via a network model of the interface. A network representation is interesting because it allows measuring local to global changes and to investigate the capacity of proteins to cope with perturbation [25]. The relevance of network models in the study of structures for protein dynamics is now well established[26-32]. The second achievement is the AAR algorithm which quantifies all structural changes between wild-type and mutant structures by simply counting the changes in their number of atomic interactions. AAR is fast (less than one second for a protein of 103 amino acids), thorough and applicable on the Cartesian coordinates of any atomic structures.

One novel finding is that structural changes follow a cascade mechanism where the local reorganization of the atoms at the site of the mutation disturbs the chemical neighbors of the mutated residue which in turn disturb their chemical neighbors, etc as in a domino effect. What triggers the cascade is not yet identified but it is neither the degree nor the weight of the original residues nor the fraction of local changes. This differs from networks where perturbations propagate through hubs (highly connected nodes) [33]. Instead, the changes propagate stepwise from hotspot to hotspot, from the site of the mutation to its neighbors (local change) to the rest of the protein (global change). This cascade mechanism results in major changes in interactions stretching out to large distances, or to more subtle changes. As mentioned already, the formers are consistent with pathological mutations while the latters accommodate adaptability and emergence of new functions through structural rearrangements which do not completely modify the protein conformation[7]. A cascade mechanism is also consistent with allostery, although multiple perturbations -as found in binding- are not tested here [34]. The cascade mechanism is more reliable than propagation of changes through hubs in a network with a power law distribution (few hubs, many low degree) because it tallies with experimental evidences on the functional impact of mutations. In a hub-regulated network, the mutation of hubs would lead to massive change, and pathologies; the mutation of residues with low degree would lead to local changes and explain robustness[35, 36]. Yet, it would be difficult to account for the emergence of new function through few subtle changes as well as for adaptive mutations (non-additive mutation effects), since there would be little or large changes. Moreover, proteins do not have hubs in terms of having nodes with a significantly higher degree than other nodes, they have nodes with average degree [25].

The second novelty is the mechanism of robustness through alternative structures, rather than just unchanged structures. This fits the updated definition of protein function: an ensemble of conformations [37]. This also lays the ground for adaptability because it allows for non-additive effects, error corrections or epistasis [6, 38]. The presence of backup links in the WT network, which allows addition and depletion of links without altering substantially the network layout, might be a clue for identifying what triggers the cascade. Backup and alternative solutions are a current mechanism of robustness, reported for other real networks such as peer-to-peer networks or other biological networks [39, 40].

In summary, the extent of structural changes produced by mutations does not depend on the degree of the mutated residue, and it does not condition the impact of a mutation on the structure. The impact of mutation involves more complex mechanisms which remain to be deciphered[41]. Altogether the mechanisms of structural changes observed through an *in silico* approach are consistent with all known functional effects of mutations (robustness, innovation, adaptation and pathology) supporting the approach as well as the hypothesis that structural robustness is embedded in the structure of the protein.

Acknowledgements

We are very thankful to Paul Sorba and Sylvie Ricard-Blum for critical reading of the manuscript. Financial support of the Federation of Research MSIF (Modelisation, Simulations and Interaction Fundamentals) is gratefully acknowledged.

References

1. D. N. Ivankov, A. V. Finkelstein and F. A. Kondrashov, *Curr Opin Struct Biol*, 2014, **26**, 104-112.
2. A. Toth-Petroczy and D. S. Tawfik, *Curr Opin Struct Biol*, 2014, **26C**, 131-138.
3. G. Amitai, R. D. Gupta and D. S. Tawfik, *HFSP J*, 2007, **1**, 67-78.
4. E. A. Ortlund, J. T. Bridgham, M. R. Redinbo and J. W. Thornton, *Science*, 2007, **317**, 1544-1548.
5. M. L. Salverda, E. Dellus, F. A. Gorter, A. J. Debets, J. van der Oost, R. F. Hoekstra, D. S. Tawfik and J. A. de Visser, *PLoS genetics*, 2011, **7**, e1001321.
6. O. Demir, R. Baronio, F. Salehi, C. D. Wassman, L. Hall, G. W. Hatfield, R. Chamberlin, P. Kaiser, R. H. Lathrop and R. E. Amaro, *PLoS Comput Biol*, 2011, **7**, e1002238.
7. R. N. McLaughlin, Jr., F. J. Poelwijk, A. Raman, W. S. Gosal and R. Ranganathan, *Nature*, 2012, **491**, 138-142.
8. L. W. Ruddock, J. J. Coen, C. Cheesman, R. B. Freedman and T. R. Hirst, *J Biol Chem*, 1996b, **271**, 19118-19123.
9. L. W. Ruddock, S. P. Ruston, S. M. Kelly, N. C. Price, R. B. Freedman and T. R. Hirst, *J Biol Chem*, 1995, **270**, 29953-29958.
10. M. J. De Wolf, G. A. Van Dessel, A. R. Lagrou, H. J. Hilderson and W. S. Dierick, *Biochemistry*, 1987, **26**, 3799-3806.
11. J. Zrimi, A. Ng Ling, E. Giri-Rachman Arifin, G. Feverati and C. Lesieur, *PLoS One*, 2010, **5**, e15347.
12. J. Hsin, A. Arkhipov, Y. Yin, J. E. Stone and K. Schulten, *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]*, 2008, **Chapter 5**, Unit 5 7.
13. W. Humphrey, A. Dalke and K. Schulten, *Journal of molecular graphics*, 1996, **14**, 33-38, 27-38.
14. O. Bachar, D. Fischer, R. Nussinov and H. Wolfson, *Protein Eng*, 1993, **6**, 279-288.
15. M. Shatsky, R. Nussinov and H. J. Wolfson, *Methods Mol Biol*, 2008, **413**, 125-146.
16. J. Schymkowitz, J. Borg, F. Stricher, R. Nys, F. Rousseau and L. Serrano, *Nucleic acids research*, 2005, **33**, W382-W388.
17. R. Guerois, J. E. Nielsen and L. Serrano, *Journal of molecular biology*, 2002, **320**, 369-387.
18. G. Gronau, S. T. Krishnaji, M. E. Kinahan, T. Giesa, J. Y. Wong, D. L. Kaplan and M. J. Buehler, *Biomaterials*, 2012, **33**, 8240-8255.
19. M. Achoch, G. Feverati, L. Vuillon, K. Salamatian and C. Lesieur, Belgrade, Serbia, 2013.
20. C. Lesieur, *Oligomerization of Chemical and Biological Compounds*, 2014, DOI: 10.5772/58576
21. V. N. Padmanabhan, H. J. Wang and P. A. Chou, *11th Ieee International Conference on Network Protocols,*

Proceedings, 2003, DOI: Doi 10.1109/Icnp.2003.1249753, 16-27.
22. S. Boyd, A. Ghosh, B. Prabhakar and D. Shah, 2005.
23. Y. Higashimoto, Y. Asanomi, S. Takakusagi, M. S. Lewis, K. Uosaki, S. R. Durell, C. W. Anderson, E. Appella and K. Sakaguchi, *Biochemistry*, 2006, **45**, 1608-1619.
24. A. F. Poon and L. Chao, *Evolution; international journal of organic evolution*, 2006, **60**, 2032-2043.
25. L. Vuillon and C. Lesieur, *Curr Opin Struct Biol*, 2015, **31**, 1-8.
26. V. A. Feher, J. D. Durrant, A. T. Van Wart and R. E. Amaro, *Curr Opin Struct Biol*, 2014, **25**, 98-103.
27. L. Di Paola and A. Giuliani, *Current opinion in structural biology*, 2015, **31**, 43-48.
28. B. Barz, D. J. Wales and B. Strodel, *J Phys Chem B*, 2014, **118**, 1003-1011.
29. K. V. Brinda and S. Vishveshwara, *Biophys J*, 2005, **89**, 4159-4170.
30. G. Feverati, M. Achoch, L. Vuillon and C. Lesieur, *PLoS One*, 2014, **9**, e94745.
31. D. M. Leitner, S. Buchenberg, P. Brettel and G. Stock, *The Journal of chemical physics*, 2015, **142**, 075101.
32. D. M. Leitner, *The Journal of chemical physics*, 2009, **130**, 195101.
33. A. L. Barabasi and Z. N. Oltvai, *Nature reviews. Genetics*, 2004, **5**, 101-113.
34. G. M. Suel, S. W. Lockless, M. A. Wall and R. Ranganathan, *Nat Struct Biol*, 2003, **10**, 59-69.
35. Y. Y. Liu, J. J. Slotine and A. L. Barabasi, *Nature*, 2011, **473**, 167-173.
36. R. Albert, H. Jeong and A. L. Barabasi, *Nature*, 2000, **406**, 378-382.
37. G. Parisi, D. J. Zea, A. M. Monzon and C. Marino-Buslje, *Curr Opin Struct Biol*, 2015, **32C**, 58-65.
38. E. Dellus-Gur, M. Elias, E. Caselli, F. Prati, M. L. Salverda, J. A. de Visser, J. S. Fraser and D. S. Tawfik, *J Mol Biol*, 2015, DOI: 10.1016/j.jmb.2015.05.011.
39. A. Wagner, *Biophys J*, 2014, **106**, 955-965.
40. J. L. Payne and A. Wagner, *Science*, 2014, **343**, 875-877.
41. T. Liu, S. T. Whitten and V. J. Hilser, *Proc Natl Acad Sci U S A*, 2007, **104**, 4347-4352.

Figure legend.

Figure 1. Schematic of the cascade mechanism underlying the structural changes associated with mutations. As the most disruptive mutation, K69N is chosen to illustrate the paths of the structural changes going from the site of mutation to elsewhere in the interface. The K69N mutation modified the atomic interactions of twenty-two hotspots of the interface covering a distance of fifteen Ångströms. The paths of changes are schematically described by arrows going from hotspots (nodes, black circles) chemically linked to hotspots chemically linked, the chemical distances (5 Å) are illustrated by dotted semi-circles. However, because the structure is a three dimensional object, the Euclidian distance between the site of mutation and the residue modified the further from it cannot be calculated from the schematic. The geodesic distances are the number of chemical links crossed to go from one hotspot to another. The structural changes of K69N cover three chemical links.

Figure 2. Local degrees and global changes. A. Spheres of influence. Only two adjacent chains D and E of $CtxB_5$ are represented in pale and dark grey strands, respectively (PDB 1EEI). The toxin interface is in ribbon. The residues modified by mutations are spacefilled and the mutated residues are red. The left panel shows the location of the four mutated hotspots K69, A64, L31 and I39 on the WT structure. The other panels on the right are their respective spheres of influence as shown on their respective X-ray structures. **B. Weak correlation between the original weighted degree of the mutated residue and the amount of structural changes after mutation measured by AAR.** $Arank_r$ values are plotted against the weights of each hotspot -*i*- before mutation w_{iWT}. The dotted line is the linear correlation. **C. Global vs local changes.** $Arank_r$ valuesare plotted against $local_r$ values (methods, local weighted degree differences ($|w_{imut} - w_{iWT}|$). The dotted line is the linear correlation and the red line is for y = 2x.

Figure 3. Schematics of additive and non-additive mutational effects. A WT network maintaining two segments together through four links of amino acids is drawn. Two sites of mutations M_1 and M_2 are considered. **Non compensatory mutations (Upper schematic).** If M_1 implies no structural and network reorganisation, then M_2 has the same effect on the WT and M1 mutated network. **Compensatory mutations (Lower schematic).** If M_2 does not have the same effect of the WT and M_1 mutated networks, then the M_1 and WT structures and networks are different.

Figure 4. Structural robustness. A**. Networks of K69 and A64 residues, before and after mutation.** Networks of the sphere of influence with hotspots nodes and links of hotspots as links . Zoom on a subset of interfacial residues in the X-ray structures of K69 and N69 (balls and stick representation). The numbers are the sequence position of the residues. The residue 69 of chain E and the residue 67 of chain D are shown in CPK and yellow, respectively. The residues of the chain E are otherwise olored in green. The backbone shows that both structures are in the same position. **B. Networks of the spheres of influence of the residue I39 and L31, before and after mutation.** Legend as in 4A.

Figure 5. Backup network of the WT interface. Structural robustness is based on the presence of backup links that allow bearing addition and depletion of links without structural impact. The nodes of the backup networks represent the hotspots, the size of the nodes represents their degree. The links represent pairs of hotspots and the colors of the links represent the number of backup for each link within a range indicated by the color scale on the right. The reddest the link, the least backup interactions the pair of amino acid has. The arrows indicate the positions of the two nodes with weak ties (50, 31) and (53, 63). The letters on the network are the chains on which the hotspots are located**.**

Figure 6. Non-additive *in silico* mutations G334V and N345D in the p53 tetrameric domain. A. p53 WT. Left panel. The chains B (light grey) and D (dark grey) of the WT p53 are shown in backbone representation (PDB 1SAK) except for the residues of the sphere of influence of the mutation G334V, spacefilled. Right panel. As on left, but with a strand representation but in strands except for the residues indicated in balls and sticks. The cascade of changes is illustrated by arrows. **B. Networks of the WT, G334V, N345D and G334+N345D spheres of influence**. Legend as in figure 4. The mutated residues are in red. The open circles are the residues whose degrees are modified by the mutation. Arrows illustrated the path of structural changes going from the residue 334 to the residue 352. The red lines are for added (continuous) and depleted

links (dotted) of amino acids. Black thick and thin lines are for increased and decreased weights, respectively.

Table 1. Mutations features

	Global Changes					*Local changes*			
Mutations	**arank$_r$**	**# modified hotspots**	**Geodesic**	**Euclidian**	**Intra**	**$\alpha_{i\ WT}$**	**W_{iWT}**	**$\Delta\alpha_{i(Mut-WT)}$**	**$\lvert\Delta W_{i(Mut-WT)}\rvert$**
K69N	182	22	3	15	0	2	30	-1	23
R67N	178	16	4	9	0	9	101	-5	24
Y76N	143	8	2	6	0	4	40	-3	37
Q3N	120	4	1	5	0	4	43	-1	26
A64N	112	18	3	10	0	4	20	5	41
Y12N	102	10	4	9	0	4	41	-4	44
T78N	97	5	3	8	0	1	2	0	1
A32N	94	11	2	5	0	5	35	2	44
E29N	90	11	3	10	0	6	77	0	30
R73N	88	18	3	15	0	4	42	-1	32
Y27N	86	16	3	13	0	5	41	-2	8
E66N	82	15	3	13	1	2	33	0	8
A98N	80	8	2	11	0	3	23	1	38
M101N	76	11	2	13	0	6	60	0	5
F25N	72	6	2	5	0	3	39	0	28
N103K	70	7	2	5	0	4	44	-2	33
A80N	66	9	3	9	0	1	1	1	24
K23N	66	6	3	11	0	1	7	-1	7
G33N	60	7	2	6	0	3	25	1	29
T71N	56	12	10	10	0	3	31	0	4
K81N	54	5	3	9	0	1	1	0	0
D70N	53	16	5	16	1	2	28	-1	10
L77N	51	14	3	10	0	4	8	1	3
S26N	48	5	1	5	0	2	15	2	25
P2N	48	7	2	7	0	4	19	0	14
V50N	46	14	4	17	1	1	1	0	1
R35N	46	9	1	6	1	5	47	-1	9
E36N	42	15	2	14	1	5	36	-2	1
Q61N	38	10	3	8	0	4	39	0	6
A97N	38	8	2	5	1	3	34	0	15
T28N	36	8	1	5	0	4	35	3	16
E11N	36	4	2	5	0	1	15	0	12
S100N	34	5	1	5	0	2	22	1	17
T1N	33	7	1	5	0	5	32	0	1
I99N	32	8	2	10	0	3	36	1	15
P93N	32	6	2	5	0	3	31	0	3
S30N	30	7	2	5	0	5	31	2	14
I58N	26	6	2	5	0	3	10	-3	10
I74N	20	10	4	9	0	3	7	-2	5
K34N	20	4	1	5	1	3	11	0	4
L31N	18	8	1	5	0	9	74	-1	1
S60N	16	8	3	7	0	2	18	0	3
L8N	16	11	2	5	1	5	19	-2	3
K63N	16	9	3	12	0	4	19	-2	6
W88N	16	7	2	11	1	3	11	-2	7
I65N	16	8	2	11	0	1	7	0	3
M68N	16	7	2	5	0	3	31	-1	8
Q49N	16	4	1	7	1	1	7	-1	7
N4K	15	5	2	5	0	1	4	3	11
I39N	14	7	1	5	0	4	16	-1	5
P53N	12	5	2	5	0	1	3	2	3
M37N	12	4	1	5	0	3	8	-2	4
I24N	12	4	2	8	1	1	1	0	0
A102N	12	3	2	5	0	3	26	0	6
T92N	8	2	1	5	0	2	16	0	4
I96N	2	3	2	5	0	1	6	0	0
I5N	2	3	2	6	1	1	7	0	0
T47N	2	2	1	5	0	1	10	0	1

i is a hospot, k_i, its degree; W_i, its weighted degree; the Euclidian distances are Ångström.

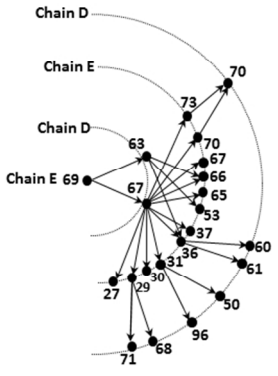


Figure 1

254x190mm (96 x 96 DPI)

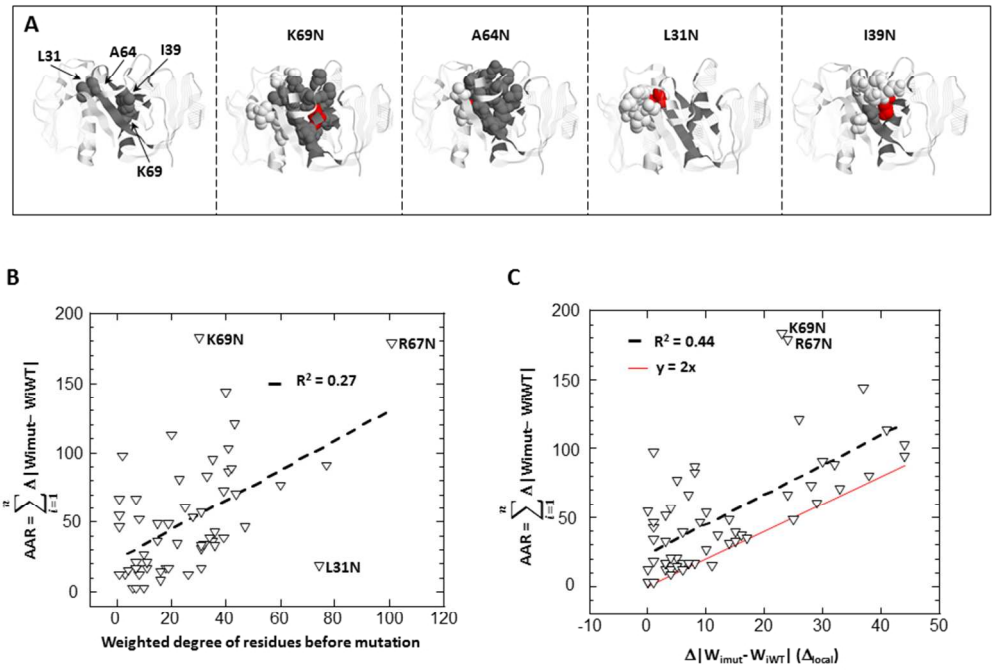


Figure 2

254x190mm (96 x 96 DPI)

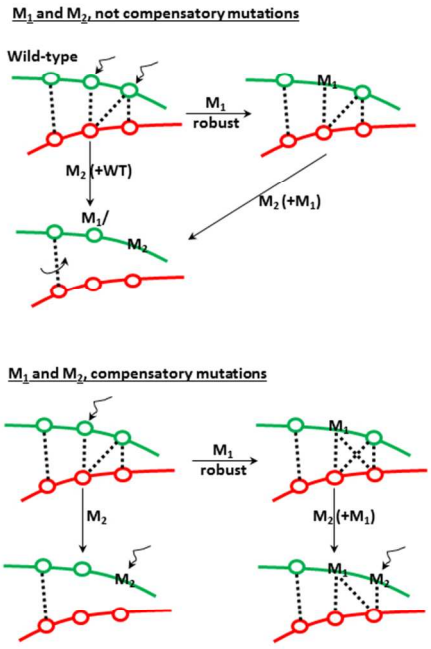


Figure 3

254x190mm (96 x 96 DPI)

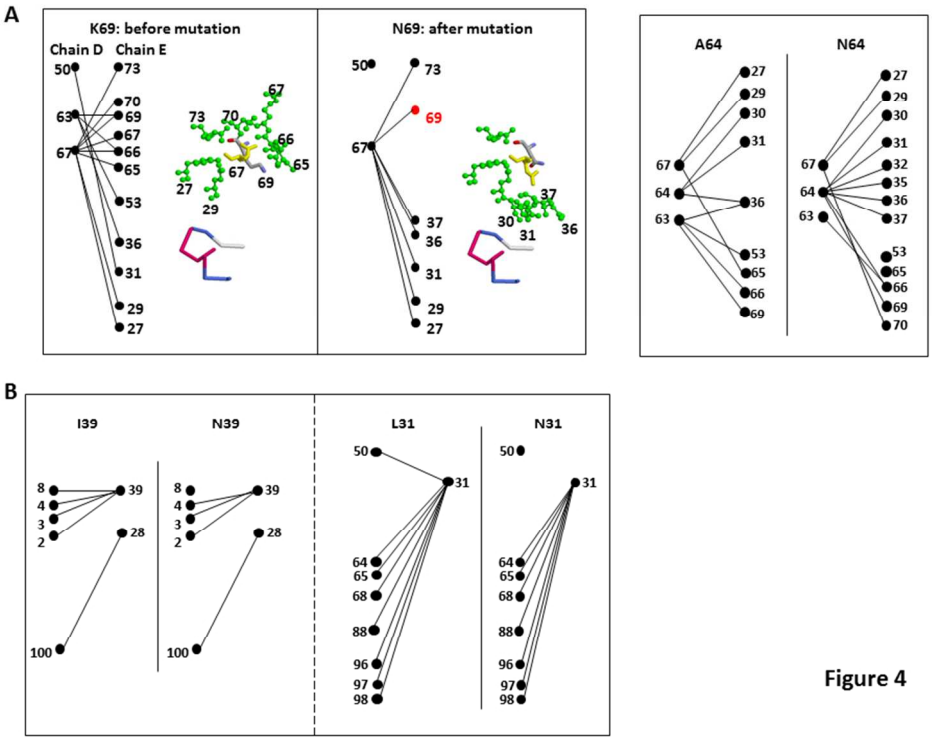


Figure 4

254x190mm (96 x 96 DPI)

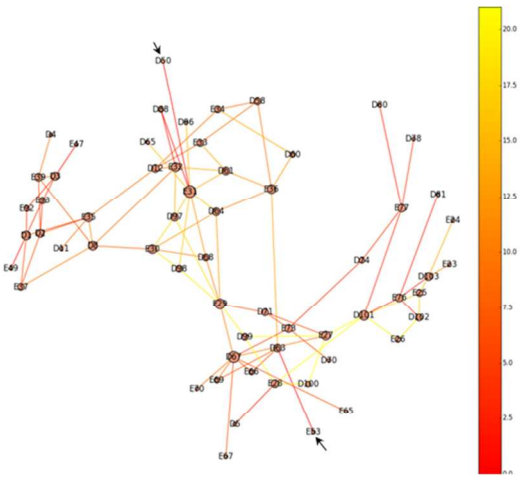
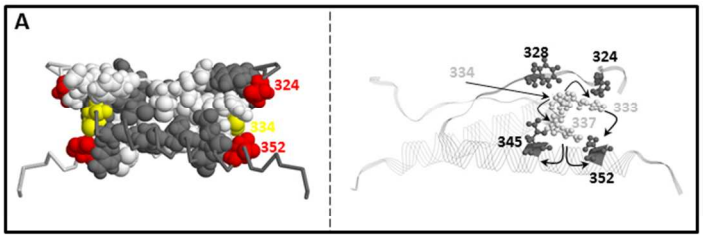


Figure 5

254x190mm (96 x 96 DPI)



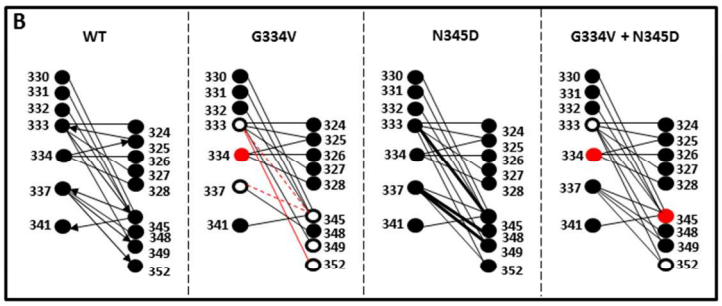


Figure 6

254x190mm (96 x 96 DPI)

Chapitre 12: Etude des propriétés de réseau protéique : Etude des quatre hots spots

Dans les chapitres précédents, nous avons montré qu'il existe des chemins de communication entre les acides aminés même s'ils sont éloignés. Puis nous avons commencé à explorer l'architecture des réseaux d'interface afin d'ouvrir des pistes pour comprendre les chemins de communication au sein des protéines. Le dernier chapitre de mon travail de thèse consiste à déterminer s'il existe des acides aminés qui introduisent une certaine faiblesse dans le réseau du fait de leur position ou de leur position et de leur nature (type d'acide aminé). Pour ce travail, j'ai sélectionné quatre hots spots de l'interface de $CtxB_5$ dont la mutation par asparagine a généré des changements structuraux et qui couvrent les propriétés chimiques et géométriques de l'ensemble des acides aminés. Il s'agit des résidus K69, R67, L31 et A98 (Figure 12.1), qui sont cette fois mutes par les 18 autres types d'acides aminés.

Cette étude a permis de classer les mutations en type d'effet locales et globales. Il s'agit maintenant d'analyser les propagations dans chaque cas pour trier les mutations robustes et les mutations fragiles pour voir si dans les quatre cas on peut trouver des types d'acides aminés robustes ou si il existe des positions supportant aucun autre type d'acide amine.

L'étude de ces quatre hots spots m'a permis de mieux comprendre le fonctionnement de l'analyse spectral, ce qui était difficile à expliquer au début de ma thèse.

Les résultats montrent que l'effet de la mutation sur l'interface a montré que la nature et l'impact de l'acide aminé n'est pas facile à analyser. Ainsi l'étude du chemin de communication est fondamental du à l'environnement et à la connectivité et que le phénomène du mécanisme en cascade existe.

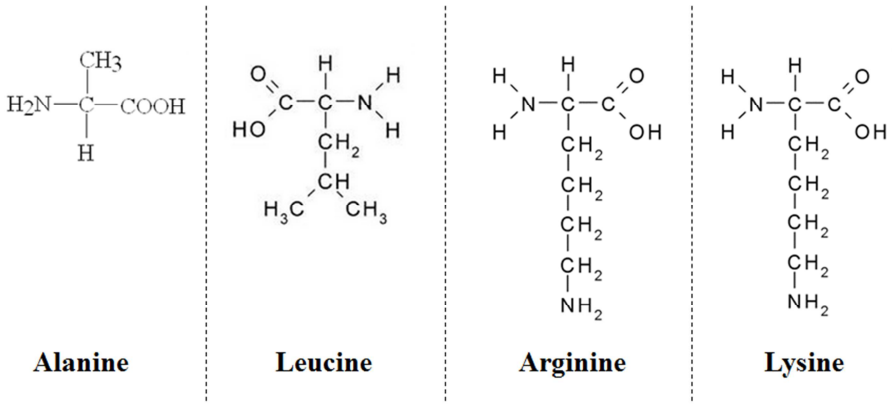


Figure 12.1 : Représentation des quatre hots spots

12.1 Protocole

Dans ce chapitre nous avons étudié les quatre hots spots (L31, A98, R67 et K69) parmi les 58 hots spots existant dans la chaîne du la toxine du choléra. Cette étude est faite par l'analyse des résultats des énergies d'interaction et la géométrie de chaque hot spot étudié en les mutant par les 19 autres acides aminés en utilisant l'algorithme Fold-X et le programme Spectral-Pro.

Pour chaque hot spot :

- Muter par les autres acides aminés : 19 fichiers de sortie mutés.
- Calculer l'énergie d'interaction pour chaque mutation de chaque hot spot.
- Calculer le nombre des interactions atome-atome (poids) et résidu-résidu (degré) pour chaque mutation de chaque hot spot.

Ces étapes nous permettent d'analyser l'hypothèse proposée, de chercher pour chaque hot spot s'il existe un ou des acide(s) aminé(s) alternatifs qui peuvent le remplacer sans perturbation de l'énergie ou du réseau.

12.2 Résultats

Rappel :

La stabilité des protéines est intimement liée à leur repliement. Les protéines ont besoin d'être dans leur état natif pour être stables. La variation d'enthalpie libre de repliement d'un état dénaturé à l'état natif s'écrit comme la contribution de l'enthalpie et de l'entropie.

$$\Delta G = \Delta H - T\,\Delta S$$

Bien que les variations d'enthalpie et d'entropie soient grandes, la variation d'enthalpie libre de l'état dénaturé à l'état natif est souvent faible (-5 à –15 kcal.mol-1). Cette énergie est comparable à celle de quelques liaisons hydrogène. La diminution d'enthalpie favorable lors du repliement est compensée par une perte d'entropie due au passage de la chaîne étendue à une structure compacte. Le terme enthalpie stabilisant la structure comprend les effets hydrophobes, les interactions de Van Der Waals et électrostatiques (en particulier les liaisons hydrogène et les ponts salins) ainsi que la formation de liaisons covalentes (les liaisons disulfure). Il est diminué par la perte des interactions qui existaient entre la structure dénaturée et le solvant (effet de désolvatation).

Les différents effets influençant la stabilité de la structure native : Les interactions qui stabilisent la structure tridimensionnelle des protéines sont principalement des interactions faibles, non covalentes : interactions électrostatiques et de Van Der Waals et effets

hydrophobes. Des liaisons covalentes sont également mises en jeu par l'intermédiaire des ponts disulfure.

Après avoir générer tous les résultats, nous présentons les énergies d'interactions et le nombre des interactions atome-atome et résidu-résidu de chaque mutation pour les quatre hots spots L31, A98, R67 et K69.

12.2.1 Hot spot A98

L'alanine (Figure 12.2) en position 98 est un hot spot existe dans la chaîne protéique de la toxine du choléra. Il est caractérisé par sa situation dans l'interface β. sa formule est la suivante :

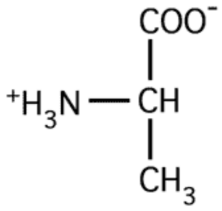


Figure 12.2 : L'alanine

Caractéristique de l'alanine

- L'alanine est un acide aminé neutre.
- Son radical se réduit à un méthyl, qui lui confère des propriétés légèrement apolaires et hydrophobes et ne permet pas de participer à des réactions chimiques.
- L'alanine représente environ 6 % des acides aminés des protéines de notre organisme.

Le tableau suivant représente les résultats obtenus des énergies d'interactions et le nombre des interactions poids et degré des 19 mutations.

Mutants	Energie d'interaction (Kcal/mol)	Em-Ewt (Kcal/mol)	poids	degré	km-kwt	1/H	poids Global	degré Global
WT	-14		23	3			1474	182
A98S	-15	-1	33	4	1	3.8	1494	184
A98M	-14	0	45	4	1	4.4	1514	184
A98G	-14	0	15	3	0	3.7	1456	182
A98L	-14	0	26	3	0	3.5	1478	182
A98V	-14	0	35	4	1	3.8	1492	184
A98C	-14	0	34	4	1	3.8	1490	184
A98I	-14	0	46	4	1	4.4	1500	180
A98T	-13	1	37	4	1	3.8	1502	184
A98Q	-13	1	63	6	3	4.8	1542	188
A98P	-12	2	38	4	1	3.9	1504	184
A98E	-12	2	62	6	3	4.6	1550	188
A98K	-10	3	67	5	2	5	1562	184
A98R	-10	3	113	8	5	6.5	1622	190
A98Y	-10	4	61	5	2	5.3	1498	184
A98N	-8	6	63	4	1	4.5	1546	184
A98D	-8	6	62	4	1	4.6	1548	184
A98F	-4	9	84	5	2	5.8	1546	184
A98H	-3	11	84	4	1	5.2	1582	182
A98W	4	18	115	6	3	6.5	1612	182

Tableau 12.1 : les résultats de calculs : énergie d'interaction, le poids et degré locaux et globaux, après la mutation de l'acide aminé alanine à la position 98 par les 19 acides aminés. Le 1/H est la mesure le nombre d'acides aminés sont modifiés après la mutation et qu'on l'appelle la déviation d'une modification uniforme (chapitre 11).

L'alanine est le deuxième plus petit des acides aminés après le Glucine, sa chaîne latérale est présentée par le groupement CH3 qui ne permet pas à ce résidu de réagir chimiquement.

D'après les énergies d'interaction et le nombre des interactions (Tableau 12.1), nous avons analysé ces résultats comme suit :

- ✓ Energie < -12 : le nombre d'interaction du mutants est égal au nombre d'interaction du WT avec apparition et disparition du résidu T28 dans les mutations S, M, G, L, V, C et I. ces derniers présentent généralement les résidus non polaires et polaires non chargé.
- ✓ Energie > -12 : les mutations affectent plus le nombre d'interaction ainsi l'énergie d'interaction, cet effet est dû à la mutation par des acides aminés aromatiques et chargé positivement ou négativement.

Les mutations de A98 par les acides aminés présentent un effet de voisinage, tel que A98 est un nœud qui contrôle la géométrie de la structure primaire, c'est-à-dire une séquence

produise en local une géométrie fiable : addition, repliement et maintien des interactions intermoléculaire en local (domaine 29-31 plus ou moins grand). Les résultats illustrés dans le tableau montrent qu'il y a une corrélation entre ΔG et les mesures du poids et degré local et global.

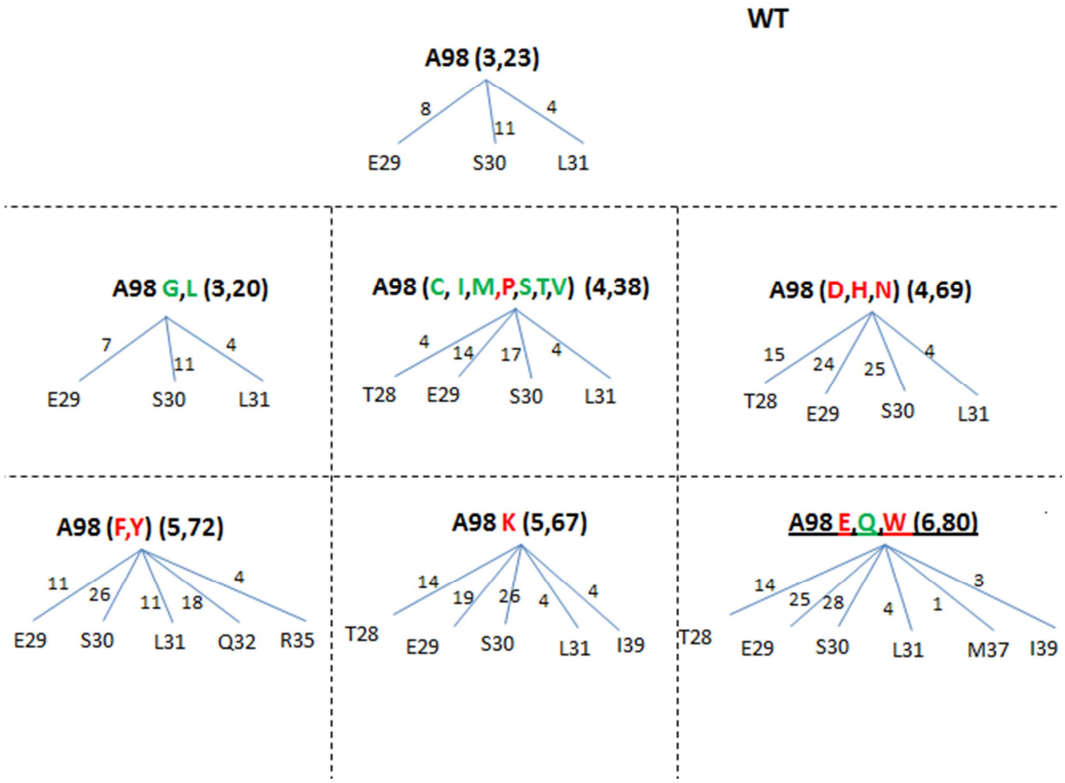


Figure 12.3 : les réseaux d'interactions de A98 pour la mutation des 19 acides aminés

Dans cette figure 12.3, les traits bleus présentent les interactions entre acides aminés. Comme nous avons observé précédemment, quand la mutation déstabilise la protéine est correspond à l'ajout des liens, par exemple la mutation par l'acide aminé glutamine (Q), il a une énergie de -13 Kcal/mol et beaucoup de liens par rapport au WT. Même l'alanine est un petit acide aminé, la plupart des mutations ont augmenté le degré car les acides aminés remplaçant ont plus d'atomes.

On remarque qu'il n y a pas de corrélations strictes entre nombre d'atomes et nombre de liens (exemples l'acide aminé leucine a le même nombre de lien que le WT mais il a plus de nombre d'atome), le degré d'un acide aminé est plus complexe que sa capacité, l'adaptation de la géométrie de l'acide aminé et sa capacité à son environnement (résidus) est aussi important.

Les résultats du calcul de la déviation d'une modification uniforme (mesure combien d'acides aminés sont modifiés après la mutation) restent constants indiquant une certaine robustesse à la position certainement dû à l'environnement qui reste le même.

12.2.2 Hot spot L31

La Leucine (Figure 12.4) en position 31 est un acide aminé parmi les hots spots existant dans la chaîne protéique de la toxine du choléra. Il est caractérisé par sa situation dans l'interface B. sa formule est la suivante :

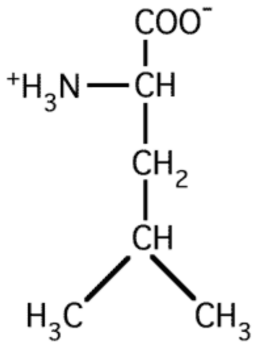


Figure 12.4 : Acide aminé Leucine

Caractéristique de la leucine

- La leucine est un acide aminé branché (6 carbones).
- Le radical de la leucine est un carbure saturé et branché, mais symétrique. Il est très hydrophobe.
- L'hydrophobie du radical de la leucine permet la formation de liaisons faibles (dites liaisons hydrophobes) avec d'autres acides aminés hydrophobes (Ile, Phe, ...) qui contribuent à la structure tertiaire et quaternaire des protéines.
- La leucine représente environ 8 % des acides aminés des protéines de notre organisme.

Pour savoir le rôle de cet acide aminé, nous avons traité tous les cas possibles de sa mutation par 19 autres acides aminés. A cette fin, nous avons calculé l'énergie d'interaction ainsi que le nombre des interactions, leur poids des liens et degré des nœuds. Les résultats sont présentés dans le tableau 12.2 suivant :

Mutants	Energie d'interaction (Kcal/mol)	Em-Ewt (Kcal/mol)	poids	degré	Km-Kwt	1/H	poids Global	degré Global
WT	-14		76	9			1474	182
L31M	-14	0	73	9	0	3.1	1474	182
L31R	-14	0	129	13	4	5.1	1588	192
L31I	-14	0	76	8	-1	3.9	1480	180
L31V	-13	1	62	8	-1	3.6	1454	182
L31K	-12	2	97	11	2	1.9	1522	186
L31N	-12	2	73	8	-1	4.1	1476	180
L31C	-11	3	44	7	-2	6.3	1412	180
L31T	-11	3	60	7	-2	4.8	1444	180
L31Q	-11	3	89	10	1	3.5	1506	184
L31A	-11	3	37	6	-3	6.5	1396	178
L31W	-11	3	162	11	2	3.2	1652	190
L31S	-11	3	49	8	-1	4.8	1428	182
L31Y	-10	3	117	10	1	6	1554	184
L31P	-10	3	58	6	-3	5.9	1428	174
L31G	-10	4	26	5	-4	6.6	1382	176
L31E	-9	5	92	10	1	2.2	1510	184
L31D	-9	5	75	8	-1	4.2	1478	180
L31H	-5	9	101	8	-1	6.8	1508	180
L31F	-3	11	109	9	0	5.9	1538	182

Tableau 12.2 : les résultats de calculs : énergie d'interaction, le poids et degré locaux et global, après la mutation de l'acide aminé leucine à la position 31 par les 19 acides aminés.

La mutation du hot spot leucine à la position 31 par les 19 autres acides aminés se divisé en deux grandes catégories :

- La première catégorie était montrée dans l'absence de l'interaction avec le résidu Valine à la position 50 dans 50 % des 19 acides aminés tel que (G,D,H,I,N,S,V,C,T,P,A) et des interactions apparaissent dans les trois résidus E51, H57 et A95.
- La deuxième catégorie était présentée dans l'apparition de nouvelles interactions avec les résidus Q49, E51, H57 et A95 dans les mutations par les résidus E, Q, W, K et R.

J'ai remarqué par cette étude des interactions qu'il y a plus de changement en cas des mutations par des résidus chargés et aromatiques, ces changements sont dû aux caractéristiques chimiques de la chaîne latérale de chacun des acides aminés tel que l'absorption en UV pour les radicales aromatique et les réactions chimiques dans le cas des radicaux chargé positivement. Sachant que le résidu L31 contrôle la géométrie tridimensionnelle dans l'interface puisqu'il se connecte avec des résidus à longue distance sur

la chaîne de la protéine (région 50, région 61-68, région 88 et région 96-98). D'après la figure suivante (Figure 12.5), on constate que ce contrôle est perdu lors des mutations c'est-à-dire qu'il y a une catégorie qui ne maintient pas la liaison avec le V50 (le seul weak ties) et une catégorie qui la maintient et même le backup. Donc la position 31 est nécessaire pour la formation même si on met n'importe quel acide aminé, il reste toujours connecte aux mêmes régions. On peut dire qu'il n'y a pas de corrélation entre les propriétés de réseaux et la mesure de ΔG.

Le calcul de la déviation d'une modification uniforme (1/H) montre l'influence des mutations qui ont plus de liens (exemple les acides aminés W, E, et K).

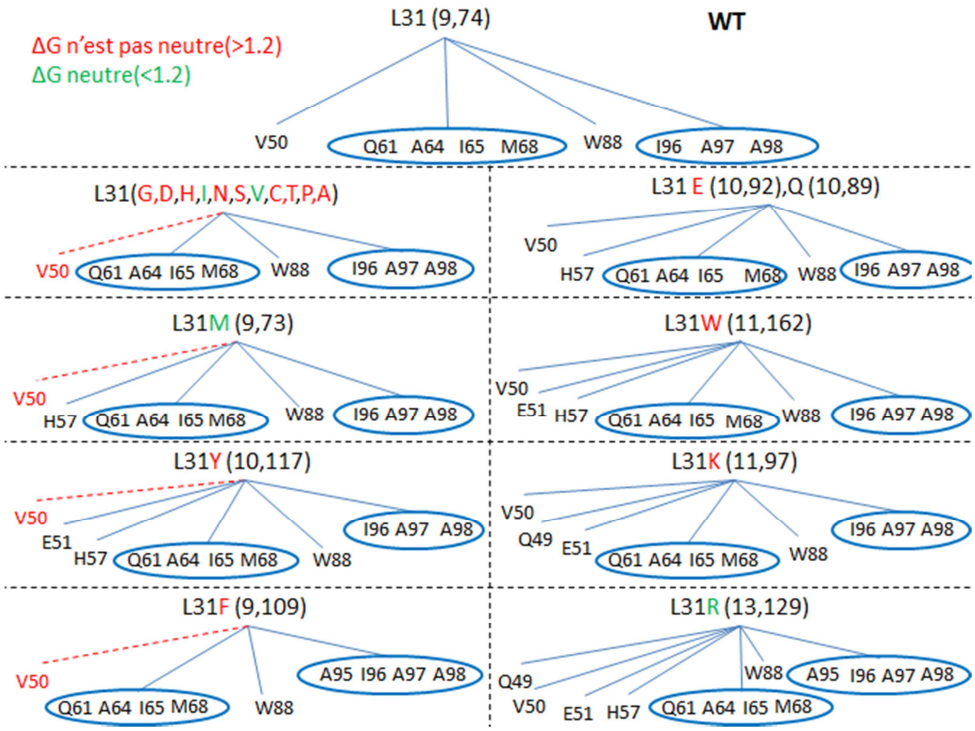


Figure 12.5 : les réseaux d'interactions de L31 pour la mutation des 19 acides aminés. Les traits bleus présentent les interactions telles que le trait rouge pointé est l'interaction disparu.

12.2.3 Hot spot K69

Le dernier hot spot étudié par sa mutation avec les 19 autres acides aminés est la Lysine (Figure 12.6). Cette étude est faite par les calculs des énergies d'interaction et le nombre d'interaction poids et degré, sa formule est la suivante :

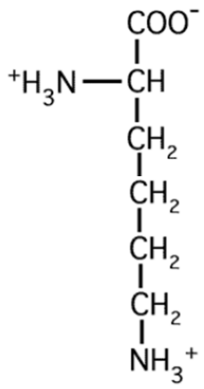


Figure 12.6 : Lysine

Caractéristique de la lysine

- La lysine est un acide aminé basique.
- Le radical comporte une chaîne de 4 carbones suivie d'une fonction amine, dont un électron ionisé crée une charge positive.
- Le radical de la lysine est très basique (pK = 10,8) ce qui lui confère une charge positive stable. De ce fait, le radical de la lysine est très polaire et très hydrophile. Cette charge positive crée des liaisons électrovalentes avec des charges électriques négatives d'autres acides aminés (Asp, Glu,...), contribuant à la structure tertiaire et quaternaire de la protéine. L'arginine entre aussi dans la structure de nombreux sites de liaison où elle contribue à fixer le ligand par des liaisons électrovalentes.
- La lysine représente environ 8 % des acides aminés des protéines de notre organisme.

Le tableau 12.3 présent les résultats obtenus des énergies d'interaction ainsi le nombre d'interaction.

Mutants	Energie d'interaction (Kcal/mol)	Em-Ewt (Kcal/mol)	poids	degré	Km-Kwt	1/H	Poids Global	Degré Global
WT	-14		29	1			1474	182
K69A	-22	-8	2	1	0	13	1408	176
K69S	-21	-7	9	1	0	5.3	1406	178
K69E	-21	-7	30	1	0	10	1442	174
K69I	-21	-7	19	1	0	10.2	1436	174
K69D	-20	-7	20	1	0	6	1440	174
K69M	-20	-6	25	1	0	7.4	1440	178
K69T	-20	-6	23	1	0	6.6	1454	176
K69N	-20	-6	7	1	0	12	1382	174
K69C	-19	-6	6	1	0	14.4	1390	176
K69P	-19	-5	3	1	0	6.6	1346	170
K69V	-18	-5	23	1	0	6.4	1452	176
K69L	-18	-4	27	1	0	4	1446	174
K69W	-18	-4	24	1	0	7.5	1458	174
K69R	-18	-4	25	1	0	4.3	1462	174
K69F	-17	-3	21	1	0	4.8	1446	176
K69Q	-17	-3	28	1	0	3.9	1450	174
K69G	-16	-3	9	1	0	6.1	1390	174
K69H	-16	-3	12	1	0	6.3	1426	172
K69Y	-15	-2	28	1	0	7	1462	174

Tableau 12.3 : les résultats de calculs : énergie d'interaction, le poids et degré locaux et globaux, après la mutation de l'acide aminé lysine à la position 69 par les 19 acides aminés.

La lysine est caractérisée par sa seule interaction le résidu R67 de la chaîne adjacente. La mutation de la lysine en position 69 stabilise la protéine mais il a gardé le même nombre d'interaction degré (1) (Figure 12.7), donc nous avons regardé plus le nombre d'interaction poids. Comme pour la mutation R67, on a une relation entre le nombre de lien et la stabilité de l'interface, la tendance est moins de lien augmente la stabilité (on a toujours moins de lien que le K69), ici cependant le contrôle se fait par le poids de l'interaction (poids) et non pas par le nombre de paires d'acides aminés. (back up interne).

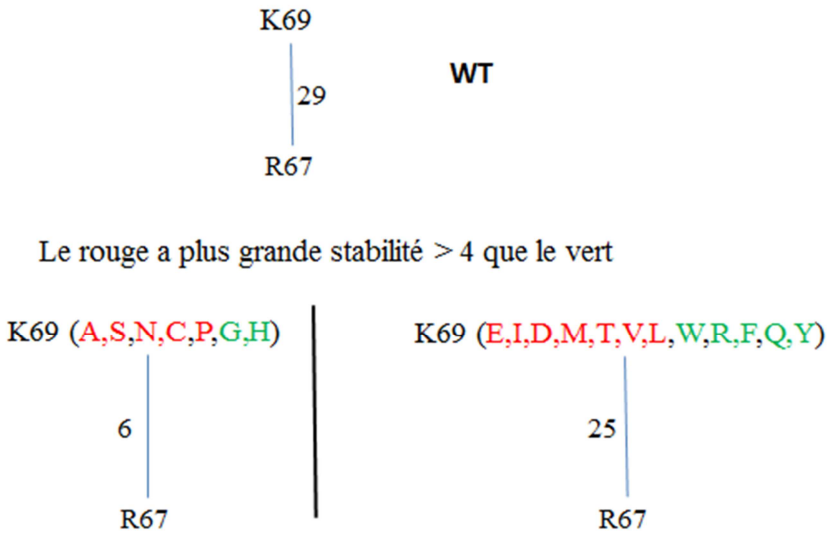


Figure 12.7 : les réseaux d'interactions de K69 pour la mutation des 19 acides aminés. Les traits bleus présentent les interactions.

En fait pour les mutations qui diminuent le poids énormément par rapport à la lysine (en jaune dans la colonne poids), il y a une forte diminution du poids global. La diminution du poids globale est moins forte pour les mutations qui maintiennent un poids local proche de celui de la lysine 69 (c'est à dire autour de 25). Par contre cet effet n'est pas vrai si on considère le degré global. Mais on ne trouve pas de lien avec les ΔG. On ne peut pas conclure qu'on ne trouve pas de corrélation entre la taille, la géométrie, la chimie et l'effet sur le ΔG. On ne voit pas non plus de corrélation avec les effets globaux. Les mesures de graphes ne sont pas les bonnes pour anticiper les effets de ΔG. On peut juste dire que la mutation préserve toujours son degré mais pas son poids. C'est donc un backup lie à des réarrangements d'atomes et pas de paires d'acides aminés.

12.2.4 Hot spot R67

Le hot spot Arginine (Figure 12.8) est un acide aminé aussi analysé dans cette étude en le mutant par les 19 autres acides aminés et calculé l'énergie d'interaction et le nombre d'interaction poids et degré, sa formule est la suivante :

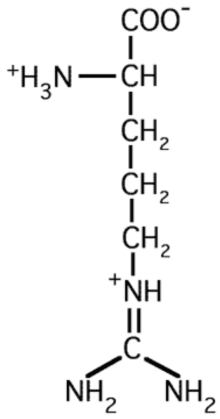


Figure 12.8 : Arginine

Caractéristique de l'arginine

- L'arginine est un acide aminé basique.
- Le radical comporte une chaîne de 3 carbones suivie d'un noyau guanidinium : un atome de carbone entouré de trois fonctions amine. Un électron ionisé crée une charge positive dans le noyau guanidinium, délocalisée dans les orbitales des trois liaisons C-N.
- Le noyau guanidinium est très basique (pK = 12,5) ce qui confère une charge positive immuable à ce radical. De ce fait, le radical de l'arginine est le plus polaire et le plus hydrophile des tous les radicaux d'acides aminés des protéines. Cette charge positive crée des liaisons électrovalences avec des charges électriques négatives d'autres acides aminés (Asp, Glu,...), contribuant à la structure tertiaire et quaternaire de la protéine. L'arginine entre aussi dans la structure de nombreux sites de liaison où elle contribue à fixer le ligand par des liaisons électrovalences.
- L'arginine représente environ 7 % des acides aminés des protéines de notre organisme.

Les résultats obtenus sont illustrés dans le tableau 12.4 suivant :

Mutants	Energie d'interaction (Kcal/mol)	Em-Ewt (Kcal/mol)	poids	degré	Km-Kwt	1/H	Poids Global	Degré Global
WT	-14		101	9			1474	182
R67F	-19	-6	78	6	-3	7.9	1392	170
R67H	-18	-5	33	4	-5	10.6	1326	172
R67W	-18	-5	63	5	-4	8.2	1408	172
R67G	-18	-4	9	2	-7	10.5	1256	166
R67Q	-18	-4	53	5	-4	9.3	1326	170
R67M	-18	-4	46	5	-4	9.1	1334	170
R67N	-18	-4	24	4	-5	9.3	1300	170
R67Y	-18	-4	65	6	-3	7.5	1384	176
R67K	-18	-4	55	6	-3	10.2	1310	174
R67S	-17	-4	24	3	-6	9.1	1284	166
R67C	-17	-4	21	2	-7	8.9	1284	164
R67E	-17	-3	22	2	-7	10.4	1270	164
R67L	-17	-3	37	4	-5	8.1	1310	166
R67I	-17	-3	21	2	-7	8.7	1280	164
R67A	-17	-3	12	2	-7	10.3	1288	166
R67V	-16	-3	22	2	-7	9.6	1292	164
R67T	-16	-3	23	2	-7	9.2	1274	162
R67P	-16	-3	15	2	-7	9.9	1248	164
R67D	-15	-1	21	2	-7	10.5	1268	164

Tableau 12.4 : les résultats de calculs : énergie d'interaction, le poids et degré locaux et global, après la mutation de l'acide aminé arginine à la position 67 par les 19 acides aminés.

L'arginine est un acide aminé chargé positivement caractérisé par ces fortes propriétés chimiques et géométriques, les résultats du calcul de nombre d'interactions montrent qu'il n'existe pas un résidu pour le remplacer. Les énergies d'interaction des mutants du hot spot R67 sont baissées l'énergie du WT c'est-à-dire la stabilisation de la protéine. Sachant que la stabilisation des protéines est due aux nombres des liens faibles entre les deux chaînes et à l'absence des liens lors de la mutation. Nous observons qu'il n'y a pas une tendance entre les énergies d'interaction et le nombre de liens (poids et degré). On voit ici que la diminution de liens semble corréler avec un abaissement du ΔG : moins de lien plus stable. On peut aussi remarquer que le contrôle de la géométrie 3D de ce nœud (il fait des paires d'acides aminés avec des résidus éloignés dans la séquence : lien avec résidus 27, 29 (une région d'interface) et lien avec résidus 65-73 (une autre région d'interface)), est maintenu dans tous les cas de mutations : on conserve toujours au moins un lien par région d'interface. On peut donc imaginer que ces réseaux sont tous redondants et équivalent à celui du WT, l'interface de la

protéine ne doit pas être métamorphosée, par contre comme il y a codifications des poids, one peut s'attendre à des changements d'affinité.

Les résultats des énergies d'interaction et du nombre des interactions atomes-atomes et résidus-résidus des 19 mutations de l'acide aminé Arginine en position 67 montrent qu'il n'y a pas de corrélation simple entre les propriétés graphes poids et degré global ou local et les ΔG.

12.3 Conclusion

La notion de chimie n'est pas suffisante pour anticiper l'effet d'une mutation : la géométrie est importante aussi. On a des géométries contraintes : difficile à reproduire : K69N : tous ces atomes en contact avec un seul résidu : le couple K69-R67 va être précieux. Backup interne n'a pas d'effet sur le degré de K69, mais des variations dans son poids.

L31 e R67: contrôle géométrie 3D, toutes les mutations affectent les pairs en intermoléculaires, il n'y a pas de contrôle interne. Addition ou déplétion ou maintien du réseau WT.

A98 : contrôle géométrie 1D : mutations modifient les voisins intermoléculaires dans un domaine 1D : 29-31 qui varient autour.

Il n'est toujours pas possible de comprendre/anticiper les modifications à ce stade.

Conclusion et discussion

Cette étude a ouvert de nouvelles pistes pour appréhender l'assemblage protéique et comprendre les mécanismes de construction des protéines, objectif biologique de mon travail de thèse. L'assemblage d'une protéine en oligomères est impliqué dans de nombreuses pathologies allant de l'infection bactérienne aux maladies de type Alzheimer ou même des cancers. Il s'agit donc d'un sujet de recherche fondamental avec une importante portée sur la santé publique. Mon approche méthodologique principale était d'étudier les propriétés des réseaux d'interactions entre atomes des acides aminés qui composent les interfaces des protéines oligomériques. Le but essentiel était de la fouille de données pour permettre d'établir si les réseaux étaient pertinents pour modéliser les phénomènes d'assemblage, quelles mesures de réseaux étaient utiles et comment les adaptés aux protéines étudiées. Pour ce travail, la ligne suivie a été la construction de réseaux d'interface et la perturbation de ces réseaux via des mutations (c'est-à-dire la modification d'un nœud du réseau). Le travail était de comprendre les propriétés des réseaux à partir des effets de perturbations. J'ai combiné l'étude du comportement d'une interface unique ($CtxB_5$) face à des perturbations à des analyses sur jeux de données.

La difficulté de cette étude de fouille de données est qu'il s'agit de construire des mesures de réseaux adéquates, c'est-à-dire répondant aux propriétés biologiques de l'objet étudié et non pas d'appliquer des « formules » réseaux toutes faites, interprétées « biologiquement » à posteriori. En bref, cette approche requiert d'appréhender les propriétés de construction d'une structure protéique, de les modéliser en termes de propriétés de réseaux et de chercher dans les données si ces propriétés existent. Ce travail se fait à partir du pentamère de la toxine du choléra, sur lesquels différents calculs de mesures sont effectués puis testés et validés ultérieurement sur des sets de données. C'est un travail minutieux et de longue haleine. La stratégie a été d'utiliser des mutations, générées par Fold-X comme source de perturbations de réseau afin de mettre en évidence les propriétés de réseaux.

De l'expérience aux réseaux (Chapitre 4, 5 et 6): la question était de trouver comment les histidines participaient à l'assemblage de la sous-unité B de la toxine du choléra, résultat montré expérimentalement, alors qu'elles n'étaient pas dans l'interface. L'approche réseau utilisée montre une régulation indirecte via des chemins de communication entre résidus, de proche en proche, suivant un mécanisme en cascade. Ainsi, la communication depuis le résidu

His 94 vers des résidus de deux régions d'interface distinctes, implique des interactions atomiques intramoléculaires avec des résidus chimiquement voisins de l'His 94 et des interactions atomiques intermoléculaires de ses voisins avec des résidus de l'interface sur la chaîne adjacente. Donc il existe une communication entre les réseaux d'interactions intramoléculaires et intermoléculaires. L'approche réseau permet d'aborder le problème de l'orchestration entre étape de repliement (intramoléculaire) et d'association (intermoléculaire). De façon intéressante, nous avons montré que le mécanisme en cascade qui permet à des résidus éloignés chimiquement de s'influencer les uns les autres s'applique aussi au réseau intermoléculaire (Chapitre 5).

Le chapitre 4 a mis en évidence des chemins de communication entre réseaux intra et inter moléculaires, chemins codés dans la séquence puisqu'il est possible de distinguer les chemins suivis par la toxine du choléra (CtxB) de ceux suivis par la toxine labile (LTB), deux toxines pourtant très proches en séquence et en structure atomique. Les deux toxines sont de plus distinguées par le nombre de leurs résidus influençant à la fois la stabilité et l'énergie d'interaction sous l'effet de mutation (Chapitre 6). Parmi ces résidus, on observe une différence topologique de réseaux. Plus de résidus communiquant entre plusieurs régions d'interface pour CtxB et plus de résidus communiquant à l'intérieur d'une région pour LTB. Il faut noter que ce résultat n'est identifié qu'à partir de mesures de perturbation, le réseau de type sauvage des deux toxines contient le même nombre de résidus impliqué dans plusieurs régions. Ce travail m'a permis de fournir une explication au fait que deux protéines de séquences et structures atomiques finales très proches s'assemblaient néanmoins de façon distinctes. Ces résultats pourront maintenant être testés d'abord sur d'autres cas du type de ces deux toxines, comme les co-chaperonnes 10 qui elles aussi partagent des hautes homologies de séquences et des structures atomiques quasi superposables, puis si ils sont validés, sur des sets de données pour établir du pronostique de mécanisme d'assemblage. Le mécanisme en cascade m'a permis aussi de proposer un cadre pour aborder la coordination des étapes d'assemblage (association et repliement) protéique. Les approches réseaux se sont donc avérées pertinentes pour répondre à mes questions.

Pour comprendre les changements de structure liés à des mutations et pouvant mener à des maladies, j'ai étudié les caractéristiques des interfaces β (deux brins β en interaction intermoléculaire) présentes dans différentes protéines et j'ai répondu à la question : est-ce que ces caractéristiques rentrent dans le mécanisme d'assemblage ? Le choix des cas et de cette

géométrie d'interface est lié aux pathologies de type Alzheimer, qui impliquent aussi des interactions entre brins β. L'idée est d'étudier les interfaces β issues de protéines 'saines' dans le sens qu'elles ne sont pas connues pour être impliquées dans des maladies de mauvaises conformations. L'hypothèse étant que contrairement aux protéines impliquées dans ce type de pathologies, les cas sains doivent avoir un mécanisme de résistance aux changements structuraux, leur permettant d'éviter par exemple la dissociation. J'ai tout d'abord étudié un petit jeu de 40 cas partageant une même symétrie, ici circulaire (Chapitre 7). Cette restriction permettait de supposer que les interfaces β étaient des précurseurs de l'assemblage et ne découlaient pas de la formation d'autres régions d'interface avant elles comme cela pourrait être le cas dans des symétries de type dihedral. Les résultats ont montré un signal insuffisant sur les acides aminés individuels des interfaces β pour pouvoir suggérer des séquences spécifiques. Nous avons pu cependant proposer un prototype de séquence de peptides potentiellement capable de reconnaitre une interface β tel que la combinatoire de séquences à tester soit nettement réduit. En effet, même si aucun motif spécifique n'a été observé dans les 40 cas, certaines positions dans le brin avaient une composition fortement réduite. En bref, une interface β était faite d'une séquence de 9-10 acides aminés dont 6-7 « hot spots » impliquant une combinatoire de 20^6 séquences (64 millions). Aux extrémités de cette séquence, on trouve préférentiellement des arginines alors qu'au centre on trouve plutôt un acide glutamique ou une histidine. De plus, le centre contient un motif alternant les acides aminés hydrophobes I, M, C et/ou W. Ces informations permettent de réduire la combinatoire à 250 séquences. J'ai pu tester expérimentalement la capacité de certaines de ces séquences à inhiber la formation de l'interface β présente dans le pentamère de la sous unité B de la toxine du choléra. Ces caractéristiques se distinguent des brins β présent dans les fibres amyloïdes ou les brins intramoléculaires par la présence de résidus chargés au centre du brin. Malheureusement, je n'ai pas pu poursuivre les tests expérimentaux pour caractériser et valider plus avant l'utilisation de ce type de peptide comme inhibiteur de fibre, faute de moyens.

J'ai donc considéré des variables de ces peptides *in silico*, ce qui m'a permis de voir des effets de non additivité et de commencer à percevoir le mécanisme d'influence existant au sein des réseaux d'acides aminés en interaction (Chapitres 4 et 5). J'ai ensuite analysé un plus grand jeu de données composé de 750 cas me permettant cette fois d'analyser les fréquences de paires d'acides aminés en plus des fréquences individuelles des acides aminés (Chapitre 8). Les résultats montrent que les fréquences des paires diffèrent significativement du produit des

fréquences individuelles suggérant un appariement ciblé. J'ai pu observer que les paires les plus fréquentes se distinguaient des paires observées dans les fibres et les feuillets β intramoléculaires, d'après la littérature. En particulier tous les feuillets β partageaient des paires communes impliquant les atomes du squelette mais pas celles impliquant des atomes des chaînes latérales. Ce résultat donne à espérer sur la possibilité de construire des inhibiteurs communs à différentes pathologies mais aussi suffisamment spécifiques pour ne pas avoir d'effets secondaires non ciblés. Là encore ces résultats n'ont pas pu être exploités expérimentalement, faute de moyens.

Cependant nous avons pu observer que l'architecture des réseaux d'interfaces β saines était différente de celle des cas pathologiques, prévenant ainsi un phénomène de propagation de dégâts structuraux, par une connectivité plus faible. Nous retrouvions de nouveau comme dans les études sur la communication intermoléculaire intramoléculaire des phénomènes non additifs et des systèmes d'influence, un processus en cascade propageant un dégât structural local dans tout le réseau. Plus en détails, les interfaces β de ces cas de protéines « saines » sont organisées en réseau « déconnecté » qui tolère la mutation d'acides aminés en bornant les effets de la mutation grâce à une faible connectivité. Nous avons proposé l'hypothèse que cette faible connectivité de réseau permettait potentiellement d'éviter la dissociation de la chaîne protéique en cas de mutation, étape préliminaire à la formation des fibres pathologiques observées dans les maladies de type Alzheimer. Les principaux résultats de cette étude indiquent que peu d'information est accessible à partir des acides aminés individuels et ce sont les paires d'acides aminés qui doivent être considérés pour aborder le problème de la construction d'une interface. De plus, la géométrie des chaînes latérales d'acides aminés est un paramètre clé pour comprendre les paires d'acides aminés.

Ces travaux ont donc montré qu'il était important de considérer les paires d'acides aminés en interaction et qu'il existait des chemins de communication au sein des structures protéiques via des systèmes influence entre paires d'acides aminés de proche en proche. Pour poursuivre ces résultats, la dernière question que j'ai abordée dans le cadre de ma thèse était d'établir les paramètres sous-jacents à cette communication. Pour y répondre j'ai étudié l'architecture des réseaux (Chapitre 9) et testé la capacité du réseau à résister à des perturbations (Chapitre 10 et 11).

Ce travail a été très laborieux puisque j'ai exploré différentes mesures qui pouvaient expliquer le facteur responsable de la propagation sans succès. Par exemple, je me suis énormément investie sur les mesures d'énergie de Fold X après mutations en tentant de les corréler à des mesures de réseaux, en vain. En regardant les interfaces générer par Fold X après mutation, j'ai réalisé que les interfaces étaient identiques avant et après mutation, car le programme ne tient pas compte de la géométrie des interactions, contrairement à Spectral-pro et Gemini. Il n'était donc pas possible de trouver une corrélation puisque les algorithmes ne mesuraient pas la même chose. Je me suis alors orientée vers une approche réseau et ai creusée la piste du rôle des degrés des nœuds dans la propagation. Le degré est connu pour être important dans le contrôle de la communication au sein de nombreux réseaux réels, les nœuds de haut degré (HUB, ou centre) servent souvent de clé de communication (loi de puissance, effet petit monde), par exemple dans les réseaux sociaux. Mais les architectures des réseaux d'acides aminés des 40 cas étudiés dans le chapitre 7, montrent une organisation en réseau avec échelle, sans nœud de haut degré type hub, et des diamètres faibles (Chapitre 9). La caractéristique petit monde est difficilement testable dans nos réseaux qui n'ont que peu de nœuds et sont donc loin d'un comportement asymptotique, condition nécessaire pour établir l'organisation de réseaux de types petits mondes. Dans l'ensemble, les architectures des réseaux ne semblent pas pointer vers une organisation ou les nœuds de hauts degrés régulent la communication au sein du réseau. De plus du fait que les contacts entre acides aminés sont limités par une surface de contacts, il n'y a qu'un facteur dix entre les acides aminés de haut degré et les acides de degré un. Ce rapport est insuffisant pour que des hubs contrôlent la propagation. Afin de tester plus avant le rôle des degrés, j'ai effectué une étude du sous-graphe de l'interface de la toxine CtxB, pour voir si la mutation d'un nœud de haut degré était plus dommageable que celle d'un nœud de bas degré (Chapitre 10 et 11). Il s'avère que le réseau est moins vulnérable au changement de degré que les réseaux sociaux par exemple, et que la mutation de nœuds de lien unique peut être plus dommageable que celle d'un nœud de haut degré. Le degré n'est pas important car les mutations mènent à des phénomènes d'influence, c'est-à-dire un mécanisme en cascade engendrant des faiblesses à longue distance. La fragilité du réseau ne se résume donc pas à un effet de connectivité du nœud mais résulte de phénomènes plus complexes (Chapitre 11).

Pour terminer, j'ai étudié profondément les propriétés de réseau protéique en me focalisant sur l'étude de quatre hots spots en particulier (Chapitre 12). Cette étude encore en cours d'analyse nous permet de cerner certains des paramètres impliqués dans les chemins de

communication (propagation). Il s'avère que la propagation est liée à une position, un type de résidu et à son voisinage. Il est intéressant de noter que le résidu K69 qui a seulement deux contacts est très sensible à la mutation alors que le résidu R67 qui a 9 contacts l'est moins. On aurait pu s'attendre à ce que n'importe quel résidu puisse remplacer le K69 qui n'a besoin que d'un atome pour faire son contact unique. Ce résultat quelque peu contre-intuitif indique simplement que pour un résidu grand comme la lysine il est très difficile de n'être contraint qu'à faire un seul contact. Ici nous voyons poindre des pistes suggérant fortement le rôle de la géométrie des voisins dans la détermination du degré d'un acide aminé. Dans les perspectives il faudra explorer le rôle du voisinage pour comprendre la propagation et l'impact de mutation. Par exemple la piste du concept de liens 'backup' qui tient compte des paires voisines comme soutient en cas de perturbations locales devrait être creusée.

Ma thèse est basée sur l'utilisation des méthodes informatiques pour répondre à un problème biologique. D'après les résultats obtenus, j'ai montré que l'approche réseau est un outil très efficace pour analyser les hypothèses sur le mécanisme d'assemblage protéique. Pour poursuivre ces travaux, j'envisagerais différentes pistes, que je n'ai pas eu le temps d'explorer.

Le premier résultat significatif et intéressant (étude de l'histidine 94, chapitre 4, 5 et 6) pour comprendre le problème biologique de l'assemblage, devrait être complété car la toxine possède trois autres histidines. Pour avoir une vision globale du problème les trois autres histidines doivent être étudiées de la même manière que l'histidine 94. La communication en cascade à l'intérieur de l'interface (réseau intermoléculaire) a été étudiée pour une région de l'interface β de la toxine du choléra. J'aurais souhaité considérer les autres régions, ainsi que d'autres protéines pour pouvoir tester si le mécanisme en cascade (système d'influence) était un système de communication général. Pour faire suite aux chapitres 11 et 12 il serait aussi important de réaliser les mutations de chaque hot spot par les 18 autres types d'acides aminés. Je n'ai pas pu réaliser ce travail du fait de la grande quantité de résultats obtenus par ailleurs et du temps nécessaire pour leur analyse.

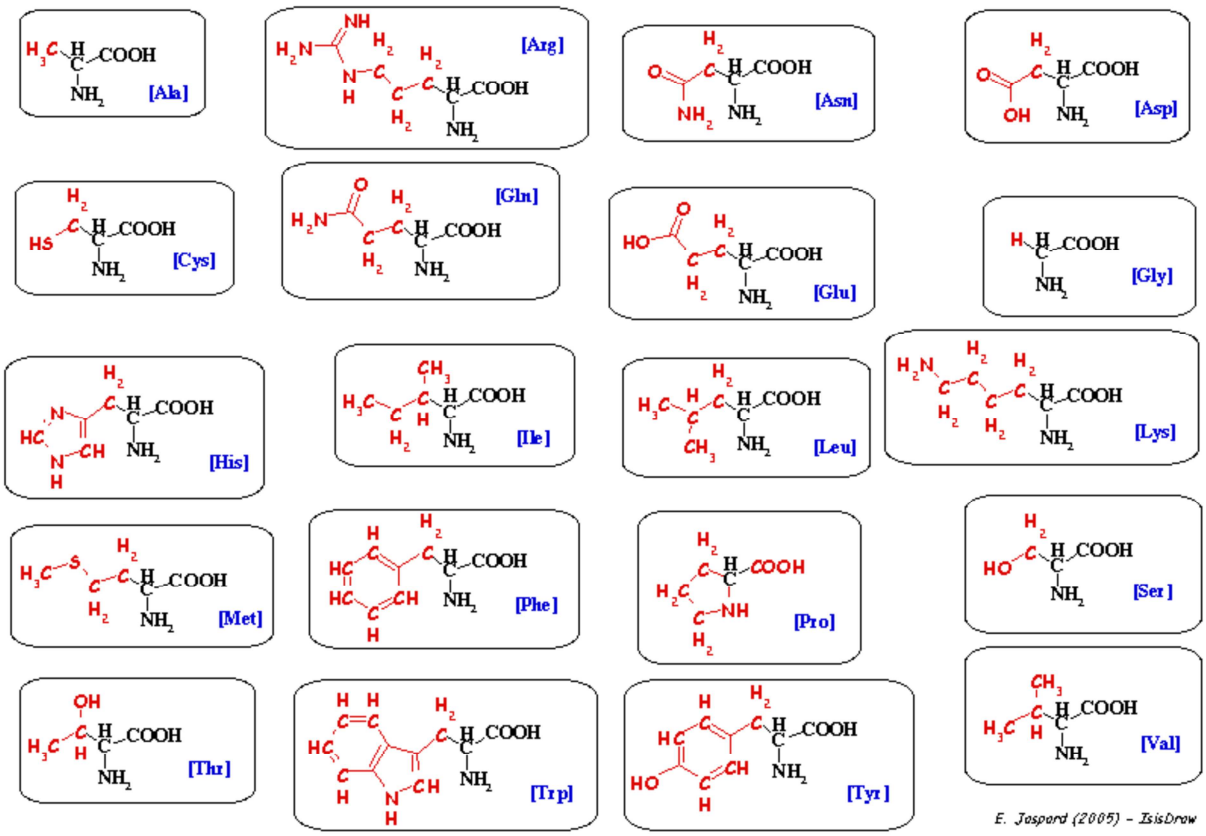
Les méthodes utilisées m'ont permis d'obtenir des résultats sur les énergies d'interactions et les propriétés de réseaux, mais je n'ai pas réussi à trouver une relation entre ces données. Il est possible que l'absence de paramètre chimique dans la construction de mon réseau en soit responsable, puisque Fold-X mesure des énergies basées sur des paramètres

chimiques uniquement. C'est-à-dire que je mesure des distances assumées être des interactions chimiques, cependant je ne regarde pas la nature chimique des atomes appariés en termes de proximité. Comme Fold-X calcule des interactions basées sur la chimie et non pas sur la distance, les deux algorithmes ne produisent pas nécessairement les mêmes résultats. Pour vérifier cette possibilité, j'ai regardé les hots spots détectés par Fold-X pour faire le calcul d'énergie d'interaction dans les mutations de hots spots (chapitre 11) et j'ai pu observer que les interfaces de Fold-X ne variaient pas en termes de hots spots avec les mutations. Pourtant Spectral-Pro montre des différences significatives dans les réseaux des mutants. Le calcul repose donc essentiellement sur des différences chimiques entre les résidus ce qui n'est pas incorporé dans Spectral-Pro.

Il faudrait peut-être introduire les hydrogènes et sélectionner la chimie par le biais de différents cut off de distances pour affiner les réseaux par type d'interaction chimique. Alternativement, comprendre les résultats de Spectral-Pro avec des mesures de réseaux tenant compte de la chimie pour comprendre la relation entre la chimie/la géométrie via des comparaisons de réseaux.

Annexe 1

Tableau des acides aminés



Annexe 2

<TITLE>FOLDX_runscript;

<JOBSTART>#;

<PDBS>;

<BATCH>list.txt;

<COMMANDS>FOLDX_commandfile;

<RepairPDB>;

<END>#;

<OPTIONS>FOLDX_optionfile;

<Temperature>298;

<R>;

<pH>7;

<IonStrength>0.050;

<water>-CRYSTAL;

<metal>-CRYSTAL;

<VdWDesign>2;

<OutPDB>true;

<pdb_hydrogens>false;

<END>#;

<JOBEND>#;

<ENDFILE>#;

Annexe 3

(a)
<COMMANDS>FOLDX_commandfile;
<RepairPDB>,Fixed;K 39;
<END>#;

(b)
<COMMANDS>FOLDX_commandfile;
<Stability>, Stability.txt;
<END>#;

(c)
<COMMANDS>FOLDX_commandfile;
<AnalyseComplex>,A;
<END>#;

(d)
<COMMANDS>FOLDX_commandfile;
<PositionScan>,KA39a,WA41d;
<END>#;

(e)
<COMMANDS>FOLDX_commandfile;
<BuildModel>, mutant_file.txt;
<END>#;

(f)
<COMMANDS>FOLDX_commandfile;
<BuildModel>, individual_list.txt;
<END>#;

Annexe 4

Chemins de communication $CtxB_5$

H94+T47+Q3	
Mutants	**intermoléculaire**
H94N	-10,9
T47N	-10,6
Q3N	-11,12
H94N+Q3N	-10,99
H94N+T47N	-10,62
T47N+Q3N	-10,78
H94N+T47N+Q3N	-10,73

H94+W88+A32	
Mutants	**intermoléculaire**
H94N	-10,9
W88N	-10,77
A32N	-9,84
H94N+W88N	-10,82
H94N+A32N	-9,98
W88N+A32N	-9,86
H94N+W88N+A32N	-10,24

H94+W88+L31	
Mutants	**intermoléculaire**
H94N	-10,9
W88N	-10,77
L31N	-8,4
H94N+L31N	-8,8
H94N+W88N	-10,82
W88N+L31N	-9,09
H94N+W88N+L31N	-10,75

H94+T92+T1	
Mutants	**intermoléculaire**
H94N	-10,9
T92N	-11,09
T1N	-11,0
H94N+T92N	-11,7
H94N+T1N	-11,1
T92N+T1N	-13,31
H94N+T92N+T1N	-14,03

H94+P93+P2	
Mutants	**intermoléculaire**
H94N	-10,9
P93N	-8,25
P2N	-11,78
H94N+P2N	-12,44
H94N+P93N	-9,02
P93N+P2N	-9,97
H94N+P93N+P2N	-10,15

H94+T92+Q3	
Mutants	**intermoléculaire**
H94N	-10,9
T92N	-11,09
Q3N	-11,12
H94N+Q3N	-10,99
H94N+T92N	-11,7
T92N+Q3N	-11,01
H94N+T92N+Q3N	-11,07

H94+Q3+P93	
Mutants	**intermoléculaire**
H94N	-10,9
Q3N	-11,12
P93N	-8,25
H94N+Q3N	-10,99
H94N+P93N	-9,02
P93N+Q3N	-8,29
H94N+P93N+Q3N	-8,41

H94+P93+T1	
Mutants	**intermoléculaire**
H94N	-10,9
P93N	-8,25
T1N	-11,0
H94N+T1N	-11,1
H94N+P93N	-9,02
P93N+T1N	-10
H94N+P93N+T1N	-13,69

Annexe 5

Cluster 1																																			
	Contacts	Liens		Contacts	Liens		Contacts	Liens		Contacts	Liens		Contacts	Liens		Contacts	Liens		Contacts	Liens		Contacts	Liens		Contacts	Liens		Contacts	Liens		Contacts	Liens		Contacts	Liens
			'E|23'	1	7																														
			'E|24'	1	1																														
'D|25'	3	13	'E|25'	2	30	'F|25'	3	15	'G|25'	3	14	'H|25'	3	18																					
'D|26'	2	12	'E|26'	2	11	'F|26'	2	11	'G|26'	2	11	'H|26'	2	11																					
'D|27'	2	24	'E|27'	4	31	'F|27'	4	25	'G|27'	4	26	'H|27'	4	28																					
'D|28'	4	29	'E|28'	4	27	'F|28'	4	30	'G|28'	4	28	'H|28'	4	31																					
'D|29'	5	60	'E|29'	6	66	'F|29'	5	63	'G|29'	5	61	'H|29'	5	59																					
			'E|53'																																
												'H|62'																							
'D|66'	2	23	'E|66'	2	28	'F|66'	2	26	'G|66'	2	24	'H|66'	2	27																					
'D|69'	1	24	'E|69'	2	26	'F|69'	1	21	'G|69'	1	17	'H|69'	1	21																					
'D|70'	2	25	'E|70'	2	27	'F|70'	2	24	'G|70'	2	23	'H|70'	2	29																					
'D|73'	4	34	'E|73'	4	32	'F|73'	5	33	'G|73'	4	33	'H|73'	4	31																					
'D|76'	4	32	'E|76'	4	28	'F|76'	4	34	'G|76'	4	28	'H|76'	4	32																					
'D|77'	5	15	'E|77'	4	8	'F|77'	4	11	'G|77'	5	19	'H|77'	4	7																					
Moyenne	3	26		3	29		3	27		3	26		3	27																					
Ecart type	1	13		1	15		1	14		1	13		1	14																					

Cluster 2																															
	Contacts	Liens		Contacts	Liens		Contacts	Liens		Contacts	Liens		Contacts	Liens		Contacts	Liens		Contacts	Liens		Contacts	Liens		Contacts	Liens		Contacts	Liens
'D|100'	2	17	'E|100'	5	30	'F|100'	5	28	'G|100'	5	26	'H|100'	5	31																	
'D|101'	5	44	'E|101'	2	17	'F|101'	2	17	'G|101'	2	17	'H|101'	2	17																	
'D|102'	3	20	'E|102'	7	48	4	6	49	'G|102'	6	47	'H|102'	6	52																	
'D|103'	4	34	'E|103'	3	20	'F|103'	3	14	'G|103'	3	19	'H|103'	3	20																	
'D|5'	1	7	'E|5'	2	10	'F|5'	1	7	'G|5'	1	10	'H|5'	1	8																	
'D|63'	4	17	'E|63'	1	8	'F|63'	1	10	'G|63'	2	10	'H|63'	1	5																	
'D|67'	9	82	'E|67'	8	76	'F|67'	7	66	'G|67'	7	80	'H|67'	7	79																	
'D|71'	3	25	'E|71'	3	22	'F|71'	3	24	'G|71'	3	25	'H|71'	2	22																	
'D|74'	3	7	'E|74'	2	6	'F|74'	2	8	'G|74'	2	5	'H|74'	3	9																	
						'F|75'						'H|75'																			
'D|78'	1	2	'E|78'	1	6	'F|78'	1	6	'G|78'	1	2	'H|78'	1	3																	
'D|80'	1	1	'E|80'	1	1	'F|80'	1	3	'G|80'	1	1	'H|80'	1	2																	
'D|81'	1	1	'E|81'	1	1	'F|81'	1	1	'G|81'	1	1	'H|81'	1	1																	
'D|99'	2	28	'E|99'	2	28	'F|99'	2	27	'G|99'	2	26	'H|99'	2	26																	
Moyenne	3	22		3	21		3	20		3	21		3	21																	
Ecart type	2	22		2	21		2	19		2	22		2	23																	

cluster 3																					
	Contacts	Liens		Contacts	Liens		Contacts	Liens		Contacts	Liens		Contacts	Liens		Contacts	Liens		Contacts	Liens
'D|1'	5	28	'E|1'	5	30	'F|1'	5	28	'G|1'	5	26	'H|1'	5	31						
'D|11'	1	12	'E|11'	1	11	'F|11'	1	13	'G|11'	1	12	'H|11'	1	13						
'D|12'	4	31	'E|12'	4	30	'F|12'	4	31	'G|12'	4	34	'H|12'	4	30						
'D|2'	4	17	'E|2'	4	20	'F|2'	4	20	'G|2'	4	14	'H|2'	4	16						
'D|3'	4	36	'E|3'	4	40	'F|3'	4	35	'G|3'	5	45	'H|3'	5	36						
'D|4'	1	4	'E|4'	1	5	'F|4'	1	4	'G|4'	1	3	'H|4'	1	4						
'D|50'	1	1	'E|50'	1	1	'F|50'	1	1	'G|50'	1	1	'H|50'	1	1						
'D|58'	3	9	'E|58'	3	10	'F|58'	3	11	'G|58'	3	10	'H|58'	3	12						
'D|60'	2	16	'E|60'	2	18	'F|60'	2	20	'G|60'	2	19	'H|60'	2	18						
'D|61'	4	28	'E|61'	4	27	'F|61'	4	32	'G|61'	4	25	'H|61'	4	33						
'D|64'	4	16	'E|64'	3	16	'F|64'	3	15	'G|64'	3	15	'H|64'	3	15						
'D|65'	1	7	'E|65'	2	9	'F|65'	1	12	'G|65'	1	6	'H|65'	1	10						
'D|68'	3	22	'E|68'	3	22	'F|68'	3	21	'G|68'	3	20	'H|68'	3	22						
'D|8'	4	12	'E|8'	4	12	'F|8'	4	12	'G|8'	4	12	'H|8'	4	12						
'D|88'	2	6	'E|88'	3	8	'F|88'	2	7	'G|88'	2	9	'H|88'	2	9						
'D|96'	1	6	'E|96'	1	7	'F|96'	1	6	'G|96'	1	6	'H|96'	1	6						
'D|97'	3	25	'E|97'	3	20	'F|97'	3	22	'G|97'	3	25	'H|97'	3	21						
'D|98'	3	15	'E|98'	3	15	'F|98'	3	15	'G|98'	3	15	'H|98'	3	15						
Moyenne	3	16		3	17		3	17		3	17		3	17							
cart type	1	10		1	10		1	10		1	11		1	10							

Cluster 4														
	Contacts	Liens		Contacts	Liens		Contacts	Liens		Contacts	Liens		Contacts	Liens
'D|35'	5	39	'E|35'	5	34	'F|35'	5	40	'G|35'	5	43	'H|35'	5	34
'D|37'	3	6	'E|37'	3	7	'F|37'	3	7	'G|37'	3	6	'H|37'	3	6
'D|39'	3	15	'E|39'	3	15	'F|39'	4	19	'G|39'	3	16	'H|39'	3	12
'D|41'	1	1										'H|41'	1	1
'D|47'	1	10	'E|47'	1	9	'F|47'	1	9	'G|47'	1	9	'H|47'	1	13
'D|49'	1	7	'E|49'	1	7	'F|49'	1	5	'G|49'	1	5	'H|49'	1	9
'D|92'	2	13	'E|92'	2	15	'F|92'	2	12	'G|92'	2	14	'H|92'	2	13
'D|93'	3	24	'E|93'	3	23	'F|93'	3	28	'G|93'	3	21	'H|93'	3	26
Moyenne	2	14		3	16		3	17		3	16		2	14
Ecart type	1	12		1	10		1	13		1	13		1	11

Cluster 5														
	Contacts	Liens		Contacts	Liens		Contacts	Liens		Contacts	Liens		Contacts	Liens
'D|30'	5	20	'E|30'	5	20	'F|30'	5	21	'G|30'	5	20	'H|30'	5	21
'D|31'	9	60	'E|31'	9	57	'F|31'	9	56	'G|31'	9	62	'H|31'	9	55
'D|32'	5	25	'E|32'	5	23	'F|32'	5	24	'G|32'	5	25	'H|32'	5	26
'D|33'	3	18	'E|33'	3	18	'F|33'	4	18	'G|33'	3	18	'H|33'	3	18
'D|34'	3	12	'E|34'	3	11	'F|34'	3	10	'G|34'	3	12	'H|34'	3	11
'D|36'	4	31	'E|36'	5	29	'F|36'	4	30	'G|36'	4	32	'H|36'	4	30
Moyenne	5	28		5	26		5	27		5	28		5	27
Ecart type	2	17		2	16		2	16		2	18		2	15

Références

1. Schlick., T., Molecular Modeling and Simulation: an interdisciplinary guide. Springer-Verlag,New-York, USA, 2002.
2. Lesieur, C., The Assembly of Protein Oligomers: Old Stories and New Perspectives with Graph Theory, in Oligomerization of Chemical and Biological Compounds, D.C.L. (Ed.), Editor. 2014, INTECH.
3. Rangarajan, N., P. Kulkarni, and S. Hannenhalli, Evolutionarily Conserved Network Properties of Intrinsically Disordered Proteins PLoS One, 2015. 10(5).
4. Csermely, P., et al., Structure and dynamics of molecular networks: A novel paradigm of drug discovery: A comprehensive review. Pharmacology & Therapeutics. 138(3): p. 333-408.
5. Barabasi, A.-L., R. Albert, and H. Jeong, Scale-free characteristics of random networks: the topology of the world-wide web. Physica A: Statistical Mechanics and its Applications, 2000. 281(1-4): p. 69-77.
6. Wang, K., S. Long, and P. Tian, Hierarchical Conformational Analysis of Native Lysozyme Based on Sub-Millisecond Molecular Dynamics Simulations. PLoS One. 10(6): p. e0129846.
7. Van Wart, A.T., et al., Weighted Implementation of Suboptimal Paths (WISP): An Optimized Algorithm and Tool for Dynamical Network Analysis. J Chem Theory Comput, 2014. 10(2): p. 511-517.
8. Demir, O., et al., Ensemble-based computational approach discriminates functional activity of p53 cancer and rescue mutants. PLoS Comput Biol, 2011. 7(10): p. e1002238.
9. Cukuroglu, E., et al., Hot spots in protein-protein interfaces: towards drug discovery. Prog Biophys Mol Biol 2014. 116: p. 165-173.
10. Lua, R.C.e.a., Prediction and redesign of protein-protein interactions. Prog Biophys Mol Biol 2014. 116: p. 194-202.
11. Sudha, G., R. Nussinov, and N. Srinivasan, An overview of recent advances in structural bioinformatics of protein-protein interactions and a guide to their principles. Prog Biophys Mol Biol, 2014. 116: p. 141-150.
12. P. Crozet, A.J.R., and M. Vervloet., Gas-phase molecular spectroscopy. Annual Reports on the Progress of Chemistry, 2002. Section C, 98:33-86.
13. Amitai, G., R.D. Gupta, and D.S. Tawfik, Latent evolutionary potentials under the neutral mutational drift of an enzyme. HFSP J, 2007. 1(1): p. 67-78.
14. Maddaluno, J., et al., Mémo visuel de chimie organique. 2015.
15. Lawrence, J., MODIFICATIONS DE LA LIAISON PEPTIDIQUE : N-HYDROXY-, N-ACYLOXY- ET N-ALKYLOXY-PEPTIDES. 2006, Universite Joseph-Fourier - Grenoble I.
16. Jarrold., M.F., Peptides and proteins in the vapor phase. Annual Review of Physical, 2000(Chemistry, 51:179-207).
17. Robisson, B., Méthodes de classification de réseaux d'interactions protéine-protéine et évaluations pour l'étude de la fonction, in Bioinformatique, Biologie Structurale et Génomique. 4 novembre 2013, Aix-Marseille Université.
18. Thomasson., W.A.B., Unraveling the Mystery of Protein Folding. Office of Public Affairs from the Federation of American Societies for Experimental Biology, 2003(Breakthroughts in Bioscience).

19. C. B. Anfinsen, R.R.R., W. L. Choate, J. Page, and W. R. Carroll., Studies on the gross structure, cross-linkages, and terminal sequences in ribonuclease. The Journal of Biological Chemistry, 207:201-210, 1954.
20. Wampler, J.E. http://bmbiris.bmb.uga.edu/wampler/tutorial/prot0.html. 1996.
21. GN, R. and S. V, Conformation of polypeptides and proteins. Adv Protein Chem, 1968. 23: p. 283-438.
22. Lesieur, C., et al., A kinetic model of intermediate formation during assembly of cholera toxin B-subunit pentamers. J Biol Chem, 2002. 277(19): p. 16697-704.
23. Ruddock, L.W., et al., Assembly of the B subunit pentamer of Escherichia coli thermolabile enterotoxin. Kinetics and molecular basis of rate-limiting steps in vitro. J Biol Chem, 1996b. 271(32): p. 19118-23.
24. Reimer, U., et al., Side-chain effects on peptidyl-prolyl cis/trans isomerisation. J Mol Biol, 1998. 279(2): p. 449-60.
25. Jakob, R.P.e.a., Elimination of a cis-proline-containing loop and turn optimization stabilizes a protein and accelerates its folding. J Mol Biol, 2010. 399: p. 331-346.
26. Joseph, A.P., N. Srinivasan, and A.G.d. Brevern, Cis-trans peptide variations in structurally similar proteins Amino Acids 2012. 43: p. 1369-1381.
27. Craveur, P., et al., Cis–trans isomerization of omega dihedrals in proteins. Amino Acids 2013. 45(279-289).
28. BENROS, C., Analyse et prédiction des structures tridimensionnelles locales des protéines, UNIVERSITE PARIS VII - DENIS DIDEROT.
29. Creighton., T.E., Proteins : structures and molecular properties. Freeman, 2002(6th edition).
30. Dobson., C.M., Protein folding and misfolding. Nature, 2003(426:884-890).
31. Karplus., C.M.D.a.M., The fundamentals of protein folding: bringing together theory and experiment. Current Opinion in Structural Biology, 1999(9:92-101).
32. NAVIZET, I.S.-L., MODÉLISATION ET ANALYSE DES PROPRIÉTÉSMÉCANIQUES DES PROTÉINES, in CHIMIE (Matière Condensée). 2004, UNIVERSITÉ PARIS 6.
33. BASTARD, K., Assemblage flexible de macromolécules : la théorie du champ moyen appliquée au remodelage des boucles protéiques, in Biologie et Sciences de la Nature. 14 septembre 2005 UNIVERSITÉ DE PARIS 7 - DENIS DIDEROT: Paris.
34. Parisi, G., et al., Conformational diversity and the emergence of sequence signatures during evolution. Curr Opin Struct Biol 2015. 32: p. 58-65.
35. Vendruscolo, M., et al., Small-world view of the amino acids that play a key role in protein folding. Phys Rev E Stat Nonlin Soft Matter Phys, 2002. 65(6 Pt 1): p. 061910.
36. Greene, L.H. and V.A. Higman, Uncovering network systems within protein structures. J Mol Biol, 2003. 334(4): p. 781-91.
37. Joel Janin, R.P.B.a.P.C., Protein-protein interaction and quaternary structure. Cambridge University Press, 2008.
38. Venn, J., On the Diagrammatic and Mechanical Representation of Propositions and Reasonings. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 1880.
39. Gursoy, A., O. Keskin, and R. Nussinov, Topological properties of protein interaction networks from a structural perspective. Biochemical Society Transactions, 2008. 36: p. 1398-1403.
40. Ma, B. and R. Nussinov, Trp/Met/Phe hot spots in protein-protein interactions: potential targets in drug design. Current topics in medicinal chemistry, 2007. 7(10): p. 999-1005.
41. Rackham, O.J., et al., The evolution and structure prediction of coiled coils across all genomes. J Mol Biol, 2010. 403(3): p. 480-93.

42. Tuncbag, N., et al., Predicting protein-protein interactions on a proteome scale by matching evolutionary and structural similarities at interfaces using PRISM. Nature Protocols, 2011. 6(9): p. 1341-1354.
43. Tuncbag, N., et al., A survey of available tools and web servers for analysis of protein–protein interactions and interfaces. Briefings in Bioinformatics, 2009. 10(3): p. 217.
44. Winter, C., et al., SCOPPI: a structural classification of protein-protein interfaces. Nucleic Acids Res, 2006. 34(Database issue): p. D310-4.
45. Feverati, G., et al., Intermolecular β-Strand Networks Avoid Hub Residues and Favor Low Interconnectedness: A Potential Protection Mechanism against Chain Dissociation upon Mutation. PloS one, 2014. 9(4): p. e94745.
46. M. Achoch, G.F., K. Salamatian, C. Lesieur Residue interaction networks : influential versus connected
47. Daily, M.D. and J.J. Gray, Allosteric communication occurs via networks of tertiary and quaternary motions in proteins. PLoS Comput Biol, 2009. 5(2): p. e1000293.
48. Bogan, A.A. and K.S. Thorn, Anatomy of hot spots in protein interfaces. J Mol Biol, 1998. 280(1): p. 1-9.
49. Clackson, T. and J.A. Wells, A hot spot of binding energy in a hormone-receptor interface. Science, 1995. 267(5196): p. 383-386.
50. Fischer, T.B., et al., The binding interface database (BID): a compilation of amino acid hot spots in protein interfaces. Bioinformatics, 2003. 19(11): p. 1453-4.
51. Thorn, K.S. and A.A. Bogan, ASEdb: a database of alanine mutations and their effects on the free energy of binding in protein interactions. Bioinformatics, 2001. 17: p. 284-285.
52. Guerois, R., J.E. Nielsen, and L. Serrano, Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. Journal of molecular biology, 2002. 320(2): p. 369-387.
53. Kortemme, T. and D. Baker, A simple physical model for binding energy hot spots in protein-protein complexes. . Proceedings of the National Academy of Sciences of the United States of America, 2002. 99: p. 14116-14121.
54. Krüger, D.M. and H. Gohlke, DrugScorePPI webserver: fast and accurate in silico alanine scanning for scoring protein-protein interactions. Nucleic Acids Res, 2010. 38: p. W480-486.
55. Lise, S., et al., Prediction of hot spot residues at protein-protein interfaces by combining machine learning and energy-based methods. BMC Bioinformatics, 2009. 10: p. 365.
56. Kim, D.E., D. Chivian, and D. Baker, Protein structure prediction and analysis using the Robetta server. Nucleic Acids Res 2004. 32: p. W526-531.
57. Schymkowitz, J., et al., The FoldX web server: an online force field. Nucleic Acids Research, 2005. 33(suppl 2): p. W382-W388.
58. Zhu, X. and J.C. Mitchell, KFC2: a knowledge-based hot spot prediction method based on interface solvation, atomic density, and plasticity features. Proteins, 2011. 79: p. 2671-2683.
59. Ofran, Y. and B. Rost, Protein–protein interaction hotspots carved into sequences. PLoS computational biology, 2007. 3(7): p. e119.
60. Vries, S.J.d. and A.M.J.J. Bonvin, How Proteins Get in Touch: Interface Prediction in the Study of Biomolecular Complexes. Current Protein and Peptide Science, 2008(9): p. 394-406.
61. Janin, J. and A.M. Bonvin, Protein-protein interactions Current Opinion in Structural Biology 2013. 23: p. 859-861.

62. Jones, S. and J.M. Thornton, Protein-protein interactions: a review of protein dimer structures. Prog. Biophys. Mol. Biol, 1995. 63: p. 31-65.
63. Ofran, Y. and B. Rost, Analysing six types of protein-protein interfaces. Journal of molecular biology, 2003. 325(2): p. 377-387.
64. Nooren, I.M. and J.M. Thornton, Diversity of protein-protein interactions. EMBO J, 2003. 22: p. 3486-3492.
65. Nooren, I.M. and J.M. Thornton, Structural characterisation and functional significance of transient protein-protein interactions. J. Mol. Biol., 2003. 325: p. 991-1018.
66. Lee, B. and F.M. Richards, The interpretation of protein structures: estimation of static accessibility. J Mol Biol, 1971. 55(3): p. 379-400.
67. Lins, L., A. Thomas, and R. Brasseur, Analysis of accessible surface of residues in proteins. Protein Sci., 2003. 12: p. 1406-1417.
68. Laskowski, R.A. and J.M. Thornton, Understanding the molecular machinery of genetics through 3D structures. Nature Reviews Genetics, 2008. 9(2): p. 141-151.
69. Teyra, J., et al., SCOWLP: a web-based database for detailed characterization and visualization of protein interfaces. BMC Bioinformatics, 2006. 7: p. 104.
70. Keskin, O., R. Nussinov, and A. Gursoy, PRISM: Protein-protein Interaction Prediction by Structural Matching. Methods Mol Biol, 2008.
71. Luke, K., M. Perham, and P. Wittung-Stafshede, Kinetic folding and assembly mechanisms differ for two homologous heptamers. J Mol Biol, 2006. 363(3): p. 729-42.
72. Shoemaker, B.A., J.J. Portman, and P.G. Wolynes and . Speeding molecular recognition by using the folding funnel: the fly-casting mechanism. Proc Natl Acad Sci U S A, 2000. 8868-73: p. 97(16): p.
73. Shoemaker, B.A., J.J. Portman, and P.G. Wolynes, Speeding molecular recognition by using the folding funnel: the fly-casting mechanism. Proc Natl Acad Sci U S A, 2000. 97(16): p. 8868-73.
74. Levy, Y., P.G. Wolynes, and J.N. Onuchic, Protein topology determines binding mechanism. Proc Natl Acad Sci U S A, 2004. 101(2): p. 511-6.
75. Wales, D.J., Energy landscapes: some new horizons. Curr Opin Struct Biol, 2010. 20(1): p. 3-10.
76. Dobson, C.B., M. A. Wozniak, and R. F. Itzhaki, Do infectious agents play a role in dementia? Trends Microbiol. 2003. 11 (7):312-317.
77. Aisenbrey, C., T. Borowik, R. Bystrom, M. Bokvist, F. Lindstrom, H. Misiak, M. A. Sani, and G. Grobner, How is protein aggregation in amyloidogenic diseases modulated by biological membranes? . Eur. Biophys, 2008(J. 37 (3):247-255.)).
78. Nathalie CARTIER-LACAVE, C.S., MALADIES NEURODÉGÉNÉRATIVES.
79. Chiti, F., and C. M. Dobson, Protein misfolding, functional amyloid, and human disease. Annu. Rev. Biochem, 2006(75:333-366).
80. Strogatz, S.H., Exploring complex networks. Nature, 2001. 410(268-276).
81. Wolf, T.I., Karev, G.& Koonin, E.V., Scale-free networks in biology: new insights into the fundamentals of evolution? BioEssays, 2002. 24, 105-10.
82. Network Science, Committee on Network Science for Future Army Applications, National Research Council. . 2005: The National Academies Press.
83. Salamatian, K., Internet science: A manifesto, in Communication Systems and Networks (COMSNETS), Fifth International Conference on. 2013. p. 1-5.
84. Sharan, R.e.a., Network-based prediction of protein function. Mol. Syst. Biol, 2007(3, 88).

85. C. von Mering, R.K., B. Snel, M. Cornell, S. Oliver, S. Fields, P. Bork, Comparative assessment of large-scale data sets of protein-protein interactions. Nature, 2002(417 (6887) () 399-403).
86. Chiquet, J., Reseaux biologiques in Gazette. 2011.
87. Barabasi, A.L. and R. Albert, Emergence of scaling in random networks. Science, 1999. 286(5439): p. 509-12.
88. Dokholyan, N.V., et al., Topological determinants of protein folding. Proc Natl Acad Sci U S A, 2002. 99(13): p. 8637-41.
89. Dubes, A.K.J.a.R.C., Algorithms for Clustering Data. Prentice Hall College Div, 1988.
90. Fortunato, S., Community detection in graphs. Physics Reports, 2010(486(3-5)).
91. Clara Pizzuti, a.S.E.R., Algorithms and tools for protein-protein interaction networks clustering, with a special focus on population-based stochastic methods..
92. Diestel, R., Graph Theory (Graduate Texts in Mathematics). 2005: Springer.
93. S. Boccaletti, V.L., Y. Moreno, M. Chavez et D.U., Complex networks : Structure and dynamics. Physics Reports. 2006(424(4-5):175-308).
94. P. Crucitti, V.L.e.S.P., Centrality in networks of urban streets. An Interdisciplinary Journal of Nonlinear Science, 2006. 015113.
95. Comin, M.N., Réseaux de villes et réseaux d'innovation en Europe : Structuration du système des villes européennes par les réseaux de recherches sur les technologies convergentes. 2009, Université de Paris I Sorbonne: Paris.
96. M. Newman, A.B., J. Watts The structure and dynamics of networks. 2006.
97. Ducruet, C. (2010) Les mesures locales d'un réseau.
98. A. Barrat, M.B.e.A.V., Dynamical Processes on Complex Networks. . Cambridge University Press, 2008.
99. Barabasi, A.L. and Z.N. Oltvai, Network biology: understanding the cell's functional organization. Nat Rev Genet, 2004. 5(2): p. 101-13.
100. Feverati, G. and C. Lesieur, Oligomeric Interfaces Under the Lens: Gemini. Laboratoire de physique théorique LAPTH, CNRS, UMR 5108 associé à l'Université de Savoie.
101. Kiel, C., L. Serrano, and al., A detailed thermodynamic analysis of ras/effector complex interfaces. J Mol Biol 340(5), 2004: p. 1039-58.
102. Schymkowitz, J., et al., The FoldX web server: an online force field. Nucleic acids research, 2005: p. W382-8.
103. de Vries, S.J. and A.M. Bonvin, How proteins get in touch: interface prediction in the study of biomolecular complexes. Curr Protein Pept Sci, 2008. 9(4): p. 394-406.
104. Goodsell, D.S. and A.J. Olson, Structural symmetry and protein function. Annu Rev Biophys Biomol Struct, 2000. 29: p. 105-53.
105. King, J. and N. Mykolajewycz, Bacteriophage T4 tail assembly: proteins of the sheath, core and baseplate. J Mol Biol, 1973. 75(2): p. 339-58.
106. Kirkitadze, M.D., G. Bitan, and D.B. Teplow, Paradigm shifts in Alzheimer's disease and other neurodegenerative disorders: the emerging role of oligomeric assemblies. J Neurosci Res, 2002. 69(5): p. 567-77.
107. Lencer, W.I., T.R. Hirst, and R.K. Holmes, Membrane traffic and the cellular uptake of cholera toxin. Biochim Biophys Acta, 1999. 1450(3): p. 177-90.
108. Zrimi, J., et al., Cholera toxin B subunits assemble into pentamers - proposition of a fly-casting mechanism. PLoS One, 2010. 5(12): p. e15347.
109. Hirst, T.R., Biogenesis of Cholera toxin and Related oligomeric Enterotoxins, ed. B.I. J. Moss, M. vaughan and A. t. Tu. Vol. 8. 1995, New York: M. Dekker. 123-184.
110. Holmgren, J., et al., Interaction of cholera toxin and membrane GM1 ganglioside of small intestine. Proc Natl Acad Sci U S A, 1975. 72(7): p. 2520-4.

111. Merritt, E.A., et al., Galactose-binding site in Escherichia coli thermolabile enterotoxin (LT) and cholera toxin (CT). Mol Microbiol, 1994b. 13(4): p. 745-53.
112. Sixma, T.K., et al., Refined structure of Escherichia coli thermolabile enterotoxin, a close relative of cholera toxin. J Mol Biol, 1993. 230(3): p. 890-918.
113. Zhang, R.G., et al., The three-dimensional crystal structure of cholera toxin. J Mol Biol, 1995. 251(4): p. 563-73.
114. Zhang, R.G., et al., The 2.4 A crystal structure of cholera toxin B subunit pentamer: choleragenoid. J Mol Biol, 1995. 251(4): p. 550-62.
115. Ruddock, L.W., et al., Kinetics of acid-mediated disassembly of the B subunit pentamer of Escherichia coli thermolabile enterotoxin. Molecular basis of pH stability. J Biol Chem, 1995. 270(50): p. 29953-8.
116. Lloyd W. Ruddock, J.J.F.C., Caroline Cheesman, Robert B. Freedman, and Timothy R. Hirst, Assembly of the B Subunit Pentamer of Escherichia coli Thermolabile Enterotoxin. BIOLOGICAL CHEMISTRY, 1996. 271.
117. C. Cheesman, L.W.R.a.R.B.F., The Refolding and Reassembly of Escherichia Coli Thermolabile Enterotoxin B-Subunit: Analysis of Reassembly-Competent and Reassembly-Incompetent Unfolded States. Biochemistry 2004.
118. Csermely, P., R. Palotai, and R. Nussinov, Induced fit, conformational selection and independent dynamic segments: an extended view of binding events. Trends Biochem Sci, 2010. 35(10): p. 539-46.
119. Hashimoto, K., et al., Caught in self-interaction: evolutionary and functional mechanisms of protein homooligomerization. Phys Biol, 2011. 8(3): p. 035007.
120. Shiou-Ru Tzeng, C.G.K., Protein dynamics and allostery. 2011.
121. Hoffman., F.M., Drosophila-abl and genetic redundancy in signal transduction. Trends Genet, 1991.
122. Launay, G., ETUDE THEORIQUE DES INTERACTIONS PROTEINE-PROTEINE, in Département de Biologie, Ecole Polytechnique, 91128 Palaiseau, France.
123. C. Lesieur, L.V., From tilings to fibers: bio-mathematical aspects of fold plasticity, in Oligomerization from chemical to biological compounds, D.C.L. (Ed.), Editor. 2014, INTECH.
124. Toth-Petroczy, A. and D.S. Tawfik, The robustness and innovability of protein folds. Curr Opin Struct Biol, 2014. 26C: p. 131-138.
125. Ortlund, E.A., et al., Crystal structure of an ancient protein: evolution by conformational epistasis. Science, 2007. 317(5844): p. 1544-1548.
126. Salverda, M.L., et al., Initial mutations direct alternative pathways of protein evolution. PLoS Genet, 2011. 7(3): p. e1001321.
127. Achoch, M., et al. Protein subunit association: NOT a social network. in 2nd International Conference “Theoretical Approaches to BioInformationSystems” TABIS.2013. 2013. Belgrade, Serbia: Institute of Physics, Belgrade.
128. Battiston, S., et al., DebtRank: too central to fail? Financial networks, the FED and systemic risk. Sci Rep, 2012. 2: p. 541.
129. Padmanabhan, V.N., H.J. Wang, and P.A. Chou, Resilient peer-to-peer streaming. 11th Ieee International Conference on Network Protocols, Proceedings, 2003: p. 16-27.
130. Boyd, S., et al. Gossip algorithms: Design, analysis and applications. in INFOCOM 2005. 24th Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings IEEE. 2005: IEEE.
131. Higashimoto, Y., et al., Unfolding, aggregation, and amyloid formation by the tetramerization domain from mutant p53 associated with lung cancer. Biochemistry, 2006. 45(6): p. 1608-1619.

132. Vuillon, L. and C. Lesieur, From local to global changes in proteins: a network view. Curr Opin Struct Biol, 2015. 31: p. 1-8.
133. Feher, V.A., et al., Computational approaches to mapping allosteric pathways. Curr Opin Struct Biol, 2014. 25: p. 98-103.
134. Di Paola, L. and A. Giuliani, Protein contact network topology: a natural language for allostery. Current opinion in structural biology, 2015. 31: p. 43-48.
135. Barz, B., D.J. Wales, and B. Strodel, A kinetic approach to the sequence-aggregation relationship in disease-related protein assembly. J Phys Chem B, 2014. 118(4): p. 1003-11.
136. Brinda, K.V. and S. Vishveshwara, A network representation of protein structures: implications for protein stability. Biophys J, 2005. 89(6): 4159-70.
137. Feverati, G., et al., Intermolecular beta-strand networks avoid hub residues and favor low interconnectedness: a potential protection mechanism against chain dissociation upon mutation. PLoS One, 2014. 9(4): p. e94745.
138. Leitner, D.M., et al., Vibrational energy flow in the villin headpiece subdomain: master equation simulations. J Chem Phys, 2015. 142(7): 075101.
139. Leitner, D.M., Frequency-resolved communication maps for proteins and other nanoscale materials. J Chem Phys, 2009. 130(19): 195101.
140. McLaughlin, R.N., Jr., et al., The spatial architecture of protein function and adaptation. Nature, 2012. 491(7422): p. 138-42.
141. Suel, G.M., et al., Evolutionarily conserved networks of residues mediate allosteric communication in proteins. Nat Struct Biol, 2003. 10(1): p. 59-69.
142. Cooper, A. and D.T. Dryden, Allostery without conformational change. A plausible model. Eur Biophys J, 1984. 11(2): 103-9.
143. Motlagh, H.N., et al., The ensemble nature of allostery. Nature, 2014. 508(7496): p. 331-9.
144. Halabi, N., et al., Protein sectors: evolutionary units of three-dimensional structure. Cell, 2009. 138(4): p. 774-86.
145. Reynolds, K.A., R.N. McLaughlin, and R. Ranganathan, Hot spots for allosteric regulation on protein surfaces. Cell, 2011. 147(7): p. 1564-75.
146. Parisi, G., et al., Conformational diversity and the emergence of sequence signatures during evolution. Curr Opin Struct Biol, 2015. 32C: p. 58-65.
147. Wagner, A., Mutational robustness accelerates the origin of novel RNA phenotypes through phenotypic plasticity. Biophys J, 2014. 106(4): p. 955-65.
148. Payne, J.L. and A. Wagner, The robustness and evolvability of transcription factor binding sites. Science, 2014. 343(6173): p. 875-7.
149. Dellus-Gur, E., et al., Negative epistasis and evolvability in TEM-1 beta-lactamase - The thin line between an enzyme's conformational freedom and disorder. J Mol Biol, 2015.
150. Liu, T., S.T. Whitten, and V.J. Hilser, Functional residues serve a dominant role in mediating the cooperativity of the protein ensemble. Proc Natl Acad Sci U S A, 2007. 104(11): p. 4347-52.
151. Vogelstein, B., et al., Cancer genome landscapes. Science, 2013. 339(6127): p. 1546-58.
152. Ciriello, G., et al., Emerging landscape of oncogenic signatures across human cancers. Nat Genet, 2013. 45(10): p. 1127-1133.
153. Nussinov, R. and C.J. Tsai, 'Latent drivers' expand the cancer mutational landscape. Curr Opin Struct Biol, 2015. 32C: p. 25-32.